Special Issue Paper

Statistics in Medicine

Received 9 January 2012, Accepted 11 January 2012 Published online 25 April 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5362

Simulation of massive public health data by power polynomials

Hakan Demirtas,[a*†] Donald Hedeker[a,b] and Robin J. Mermelstein[b]

Situations in which multiple outcomes and predictors of different distributional types are collected are becoming increasingly common in public health practice, and joint modeling of mixed types has been gaining popularity in recent years. Evaluation of various statistical techniques that have been developed for mixed data in simulated environments necessarily requires joint generation of multiple variables. Most massive public health data sets include different types of variables. For instance, in clustered or longitudinal designs, often multiple variables are measured or observed for each individual or at each occasion. This work is motivated by a need to jointly generate binary and possibly non-normal continuous variables. We illustrate the use of power polynomials to simulate multivariate mixed data on the basis of a real adolescent smoking study. We believe that our proposed technique for simulating such intensive data has the potential to be a handy methodological addition to public health researchers' toolkit. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Fleishman polynomials; biserial correlation; phi coefficient; simulation; tetrachoric correlation; random number generation

1. Introduction and motivation

Stochastic simulation is a consequential aspect of statistical research. Model building, estimation, and testing steps typically require verification via simulation to assess the validity, the reliability, and the plausibility of inferential techniques, to evaluate how well the implemented statistical models capture the specified population quantities, and to evaluate how reasonably these models respond to departures from underlying assumptions, among other things. Generating data in a repeated manner allows researchers to study the performance of their statistical methods through simulated data replicates that mimic the real data characteristics of interest in any given setting. The resulting accuracy and precision measures regarding the parameters under consideration signal if the procedure works properly and may suggest remedial action to minimize the discrepancies between expectation and reality.

In public health sciences, most data sets include different types of variables. This work is motivated by the need of simultaneously simulating binary and possibly non-normal continuous data given the marginal characteristics of each variable as well as the linear association structure among variables in the system. To the best of our knowledge, no methodology has appeared in the literature that addresses this problem, and we attempt to fill this gap.

We built the proposed mechanism upon a combination of a few random variate generation routines that involve simulation of correlated binary data, multivariate normal (MVN) data, a mix of binary and normal data, and multivariate non-normal continuous data. MVN data generation is well established; and the 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 [1]. A computational routine for joint generation of a binary and normal data, proposed by Demirtas and Doganay [2], forms an

[a]*Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL 60612, U.S.A.*
[b]*Institute for Health Research and Policy, University of Illinois at Chicago, Chicago, IL 60608, U.S.A.*
**Correspondence to: Hakan Demirtas, Division of Epidemiology and Biostatistics, University of Illinois at Chicago, 1603 West Taylor Street, MC 923, Chicago, IL 60612, U.S.A.*
[†]*E-mail: demirtas@uic.edu*

algorithmic basis for augmenting the continuous part so it can accommodate non-normal data and is driven by the relationship among phi coefficient, point biserial, and tetrachoric (biserial) correlations. For the non-normal continuous part, we resort to an extension of Fleishman's power polynomial 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 MVN data whose components yield the non-normal data through the coefficients of powers of normal variates [3].

The method assumes that all variables in a given data set jointly follow an MVN density after finding the normal components of continuous variables in Fleishman's system by solving a set of non-linear equations, and some of the components are then dichotomized. Two sets of correlations naturally get altered with this operation: (i) correlation among dichotomized variables and (ii) correlations among normal and dichotomized variables. The magnitude of the first change needs to be computed through integration, and that of the second change comes from a simple formula from the dichotomization literature. Once these transitions are performed, one can form an overall correlation matrix for an MVN distribution that would lead to the specified correlations after dichotomizing some of the variables via thresholds that are determined by marginal proportions. The final step is going back to the original scale of continuous variables by the aforementioned set of coefficients. As long as some conditions that we outline in Section 3 hold, this method is capable of generating data that follow the specified linear association structure for all variables, means of binary variables, and skewness and kurtosis behavior for continuous variables.

The organization of the rest of the article is as follows. In Section 2, we provide background information on how to generate 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. We obtain the dichotomized versions by the specified marginal proportions. We also discuss how the magnitude of the correlation changes when only one variable is dichotomized for bivariate data. Last but not least, we delineate in more detail the technique of power polynomials for generating multivariate continuous data. In Section 3, we outline the proposed methodology to generate mixed data. In Section 4, we present a simulation work that is devised around an adolescent smoking behavior data for evaluating the performance of the suggested approach by commonly accepted accuracy and precision measures. Section 5 includes discussion, concluding comments, and future directions.

2. Background

In this section, we give fundamental characteristics of MVN data generation, multivariate binary data generation, and dichotomization as well as univariate and multivariate Fleishman polynomials [4, 5].

2.1. *Multivariate normal data generation*

Sampling from MVN distribution is straightforward. Suppose $Z \sim N_d(\mu, \Sigma)$, where μ is the mean vector, Σ is symmetric, positive definite, and $d \times d$ is the variance–covariance matrix. The MVN density is always finite; the integral is finite as long as $|\Sigma^{-1}| > 0$. A random draw from an 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 $AA^{\mathrm{T}} = \Sigma$. If $z = (z_1, \ldots, z_d)$ 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.

2.2. *Multivariate binary data generation*

Although several multivariate Bernoulli data simulation routines appeared in the literature [6, and references therein], the one that fits into our framework was proposed by Emrich and Piedmonte [1] who introduced a method for generating correlated binary data in which the joint distribution of the binary variables is completely determined by 'borrowing' the third-order and higher-order moments from an MVN distribution. Let $Y_1, \ldots, Y_j$ represent binary variables such that $E[Y_j] = p_j$ and $\mathrm{Cor}(Y_j, Y_k) = \delta_{jk}$, where p_j $(j = 1, \ldots, d)$ and δ_{jk} $(j = 1, \ldots, d-1; k = 2, \ldots, d)$ are given and where $d \geq 2$. As Emrich and Piedmonte [1] noted, δ_{jk} is bounded below by $\max\left(-\sqrt{(p_j p_k / q_j q_k)}, -\sqrt{(q_j q_k / p_j p_k)}\right)$ and above by $\min\left(\sqrt{(p_j q_k / q_j p_k)}, \sqrt{(q_j p_k / p_j q_k}\right)$, where $q_j = 1 - p_j$. Let $\Phi[x_1, x_2, \rho]$ be

the CDF for a standard bivariate normal random variable with correlation coefficient ρ. Naturally, $\Phi[x_1, x_2, \rho] = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(z_1, z_2, \rho)\mathrm{d}z_1\mathrm{d}z_2$, where $f(z_1, z_2, \rho) = [2\pi(1-\rho^2)^{1/2}]^{-1} \times \exp\left[-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right) / \left(2(1-\rho^2)\right)\right]$. We could generate MVN outcomes (Z's) whose correlation parameters are obtained by solving the equation

$$\Phi[z(p_j), z(p_k), \rho_{jk}] = \delta_{jk}(p_j q_j p_k q_k)^{1/2} + p_j p_k \tag{1}$$

for ρ_{jk} ($j = 1, \ldots, d-1; k = 2, \ldots, d$), where $z(p)$ denotes the pth quantile of the standard normal distribution. As long as δ_{jk} satisfies the range condition mentioned earlier, 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 the positive definiteness of Σ cannot be guaranteed. To create dichotomous outcomes (Y_j) from the generated normal outcomes (Z_j), we set $Y_j = 1$ if $Z_j \leqslant z(p_j)$ and 0 otherwise for $j = 1, \ldots, d$. This produces a vector with the desired properties: $E[Y_j] = P(Y_j = 1) = P(Z_j \leqslant z(p_j)) = p_j$ and $\mathrm{Cov}(Y_j, Y_k) = P(Y_j = 1, Y_k = 1) - p_j p_k = P(Z_j \leqslant z(p_j), Z_k \leqslant z(p_k)) - p_j p_k = \Phi[z(p_j), z(p_k)\rho_{jk}] - p_j p_k = \delta_{jk}(p_j q_j p_k q_k)^{1/2}$. Therefore, $\mathrm{Cor}(Y_j, Y_k) = \mathrm{Cov}(Y_j, Y_k)/(p_j q_j p_k q_k)^{1/2} = \delta_{jk}$ by Equation (1).

Of note, points below the threshold, which are determined by the marginal proportion ($z(p_j)$), are assigned a value of 1. When the direction changes ($P(Y_j = 1) = P(Z_j \geqslant z(1-p_j))$), the correlation does not change as long as the direction is consistent for both variables. This issue will be re-visited for the binary–normal combinations in Section 2.3.

2.3. *Relationships involving 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 variable 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 [7]. 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 versus non-smoker).

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 versus 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 variable 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

$$\rho_{X_D Y} = \rho_{XY}\left(h/\sqrt{pq}\right), \tag{2}$$

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 (2) should not change with dichotomization, so high and low values of X are assigned 1 and 0, respectively. Furthermore, for consistency, the same directionality will be followed for the binary–binary cases described in Section 2.2.

In the next section, we provide a unified framework on the joint generation of binary and continuous variables. Of note, for the purposes of this work, artificial versus true dichotomies or terminology

differences such as 'biserial' versus 'tetrachoric' are unimportant. What is needed at the correlation specification step is a matrix whose elements consist of Pearson correlations for the normal–normal pairs as well as phi coefficient for the binary–binary pairs and point biserial correlation for the binary–normal pairs. Again, both types of latter associations are nothing more than special names of Pearson correlations. After performing calculations given in Equations (1) and (2), one finds an overall Pearson correlation matrix for the underlying MVN realizations before dichotomizing some variables in the system.

2.4. *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 non-normal 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 $\nu_2 \geqslant \nu_1^2 - 2$.[‡] Fleishman [4] gave high-order moment (the third and fourth moments) boundaries of the power method through an inequality; however, they were not entirely correct. Subsequently, Headrick and Sawilowsky [8] computed empirical lower bounds of kurtosis for a given value of skewness.

Assuming that $E[Y] = 0$ and $E[Y^2] = 1$, by utilizing the first 12 moments of the standard normal distribution, we can derive the following set of equations after simple but tedious algebra:

$$a = -c \tag{3}$$

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0 \tag{4}$$

$$2c(b^2 + 24bd + 105d^2 + 2) - \nu_1 = 0 \tag{5}$$

$$24\left[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)\right] - \nu_2 = 0 \tag{6}$$

Solving these equations can be accomplished by the Newton–Raphson method or any other plausible root-finding or non-linear optimization routine. Demirtas and Hedeker [9] gave a computer implementation for the Newton–Raphson algorithm for this particular setting.

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. Because $a = -c$, it comes down to solving the following equations:

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

The first derivative matrix is

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

[‡]*It is trivial to prove this by Cauchy–Schwarz inequality. Furthermore, one can show that equality condition is impossible to reach, but this is immaterial for the purposes of this work.*

where

$$g_1'(b) = 2b + 6d, g_1'(c) = 4c, g_1'(d) = 6b + 30d$$

$$g_2'(b) = 2c(2b + 24d), g_2'(c) = 2(b^2 + 24bd + 105d^2 + 2), g_2'(d) = 2c(24b + 210d)$$

$$g_3'(b) = 24(d + 2bc^2 + 28c^2d + 48d^3), g_3'(c) = 24(2c + 2b^2c + 56bcd + 282cd^2)$$

$$g_3'(d) = 24(b + 28bc^2 + 24d + 144bd^2 + 282c^2d + 900d^3)$$

Updated equations in Newton–Raphson are as follows:

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

Fleishman's method has been extended in several ways in the literature. One extension utilizes the fifth-order polynomials in the spirit of controlling for higher-order moments [10]. The other one is in regard to a multivariate version of the power method [5]. The generalizability to the multivariate case makes the polynomial method more compelling in the sense that it presents an advantage over other general distributions such as Burr family [11], Johnson family [12], Pearson family [13], and Schmeiser–Deutch system [14], whose multivariate versions are either non-existent or very formidable to specify because of mathematical and/or computational difficulties. It should be noted that some authors have criticized the power approach (e.g., [15]) on the grounds that the exact distribution was unknown and thus lacked PDF and CDF. However, Headrick and Kowalchuk [16] derived the power method's PDF and CDF in general form. For a definitive source and in-depth coverage of Fleishman polynomials, see Headrick [3].

We now focus on the multivariate extension that has a central role for the remainder of this paper [5]. The procedure for generating multivariate continuous data begins with the computation of the constants given in Equations (3), (4), (5), and (6), for each variable independently. We can formulate the multivariate case in matrix notation as shown later. 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 of one of them, $\mathbf{z}' = [1, Z, Z^2, Z^3]$; 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 non-normal variable Y is then defined as the product of these two vectors, $Y = \mathbf{w}'\mathbf{z}$. Let r_{Y_1,Y_2} be the correlation between two non-normal variables Y_1 and Y_2 corresponding to the normal variables Z_1 and Z_2. As the variables are standardized, meaning $E(Y_1) = E(Y_2) = 0$, $r_{Y_1Y_2} = E(Y_1Y_2) = E(\mathbf{w_1'z_1z_2'w_2}) = \mathbf{w_1'}\mathcal{R}\mathbf{w_2'}$, 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 $r_{Y_1Y_2}$ and $\Delta_{Z_1Z_2}$ in terms of polynomial coefficients ensues:

$$r_{Y_1Y_2} = \Delta_{Z_1Z_2}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) + \Delta_{Z_1Z_2}^2(2c_1c_2) + \Delta_{Z_1Z_2}^3(6d_1d_2). \quad (7)$$

Solving this cubic equation for $\Delta_{Z_1Z_2}$ gives the intermediate correlation between the two standard normal variables that is required for the desired post-transformation correlation $r_{Y_1Y_2}$. Clearly, we should assemble correlations for each pair of variables into a matrix of intercorrelations that will be used in MVN data generation.

After reviewing the salient features of the power approach, we describe a novel methodology for generating mixed data in the next section.

3. Proposed methodology

Finding the coefficients of powers of normal components of any continuous distribution can be performed by solving a system of non-linear equations, and employing these coefficients in determining the intermediate correlations among those normal components are explained in Section 2.4. MVN and multivariate binary generation with underlying normal distribution are well understood, and along with the mathematical connection between point biserial and tetrachoric correlations described in Sections 2.1 and 2.2, 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 non-normal variates. More specifically, algorithmic steps are given in the succeeding text. In what follows, B, N, and C stand for sets of binary, normal, and non-normal 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 type 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, 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 Section 2.2.
3. Repeat step 2 for the BC and CC parts by the approximation method proposed by Demirtas and Hedeker [17], which we detail in Section 3.1.
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. If the parameters come from a real data set, store the mean and standard deviation, which will be needed in step 16, and then center and scale the continuous variables. 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 (3)–(6) given corresponding ν_1 and ν_2 values.
7. For each CC combination, using the constants in step 6, find the intermediate correlation by solving Equation (7).
8. For each BC combination, suppose that there are two identical standard normal (N) variables, one is the normal component of the continuous variable and the other one is the underlying binary variable before dichotomization. With this setup, $\text{Cor}(\text{B}, \text{N}) = h/\sqrt{pq}$ using Equation (2), substituting $+1$ for the biserial correlation (as they are identical before dichotomization).
9. Solve for $\text{Cor}(\text{C}, \text{N})$ assuming $\text{Cor}(\text{B}, \text{C}) = \text{Cor}(\text{B}, \text{N}) * \text{Cor}(\text{C}, \text{N})$. 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, $\text{Cor}(\text{B}, \text{C})$ is specified, and $\text{Cor}(\text{B}, \text{N})$ is found in step 8.
10. Compute the intermediate correlation between C and N by Equation (7). Notice that for standard normal variables, $b = 1$ and $a = c = d = 0$. So intermediate correlation is the ratio, $\text{Cor}(\text{C}, \text{N})/(b + 3d)$, where b and d are the non-zero coefficients of the non-normal continuous variable.
11. Construct an overall correlation matrix, Σ^*, using the results from steps 5 and 7–10.
12. Check if Σ^* is positive semidefinite. If it is not, find the nearest positive semidefinite correlation matrix.
13. Generate MVN data with a mean vector of $(0, ..., 0)_{k+j}$ and a 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 normals using the corresponding (a, b, c, d) coefficients.
16. Go back to the original scale for continuous variables by reverse centering and scaling.

3.1. *Operational remarks*

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 [17] 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, negative definiteness for Σ and unfeasible correlation range can never ensue. Second, Fleishman polynomials [4] do not cover the entire skewness-elongation plane. Therefore, most but not all not (ν_1, ν_2) specifications are plausible. Third, even when no aforementioned possible complications exist, the final correlation matrix, Σ^*, is not guaranteed to be positive semidefinite. In such cases, we recommend the method of Higham [18] 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.

4. A simulation study devised around a real data set

Modern data collection procedures, such as ecological momentary assessments (EMA) and/or real-time data captures, have been developed to record the momentary events and experiences of subjects in daily life [19]. These procedures yield large numbers of observations per subject. In this article, we describe data from a natural history study of adolescent smoking using EMA. Students included in this study were either 9th or 10th grade at the baseline; 55.1% were girls and self-reported on a screening questionnaire 6–8 weeks prior to baseline that they had smoked at least one cigarette in their lifetime. Written parental consent and student assent were required for participation. A total of 461 students completed the baseline measurement wave. The study utilized a multimethod approach to assess adolescents in terms of self-report questionnaires, a week-long time/event sampling method via palmtop computers (EMA), and in-depth interviews.

Here, we focus on the EMA data. Adolescents carried the handheld computers with them at all times during a seven-consecutive-day data collection period and were trained to both respond to random prompts from the computers and to event record smoking episodes. Questions included ones about place, activity, companionship, mood, and other subjective items. The handheld computers dated and time-stamped each entry. In what follows, we used the responses obtained from the random prompts. In all, there were 14,105 random prompts obtained from the 461 students with an approximate average of 30 prompts per student (range = 7–71).

Two outcomes were considered: measures of the subject's negative and positive affects (*NA* and *PA*) at each random prompt. Both of these measures consisted of the average of several individual mood items, each rated from 1 to 10, which were identified via factor analysis. Specifically, *PA* consisted of the following items that reflected a subject's assessment of their positive mood before the prompt signal: I felt happy, I felt relaxed, I felt cheerful, I felt confident, and I felt accepted by others. Similarly, *NA* consisted of the following items assessing preprompt negative mood: I felt sad, I felt stressed, I felt angry, I felt frustrated, and I felt irritable.

In addition to these two outcomes, for the purposes of this work, we included six subject-level predictors in the system. These are *SMOKER* (an indicator of whether the student is a current smoker, coded no = 0 or yes = 1; this was determined on the basis of whether or not the subject provided at least one smoking event during the week-long data collection period), *MALE* (coded 0 = female or 1 = male), *GRADE10* (coded 0 = 9th or 1 = 10th grade), *NOVSEEK* (a measure of novelty seeking), *NEGMR* (a measure of negative mood regulation), and *ALONE* (proportion of random prompts in which subject was alone). In addition, *PA* and *NA* values were averaged across the subject's repeated observations. We call these average scores *AVGPA* and *AVGNA*. With this setup, the data set we use for the remainder of this

section has five continuous (*AVGPA*, *AVGNA*, *NOVSEEK*, *NEGMR*, and *ALONE*) and three dichotomous (*SMOKER*, *MALE*, and *GRADE10*) variables.

The primary goal of this work is the ability of generating data that closely resemble the original data trends in terms of marginal and associational behavior; therefore, instead of conducting model-based analyses (see Hedeker *et al*. [20] for an example), we concentrate on the degree of agreement between real and simulated data on some key statistical quantities.

As the data set has $n = 461$ subjects, the number of subjects for simulated data was also set to 461. We generated data using the proposed technique outlined in Section 3 and stored estimates of parameters of interest. The process was repeated 1000 times. The marginal parameters we have chosen were proportions for binary variables and mean, median, standard deviation, interquartile range, skewness, and kurtosis for continuous variables. The associational parameters that we considered were odds ratios for binary variables as well as correlations among both types of variables. Because of space limitations, we only report the average estimates across 1000 simulation replicates in Tables I–III, where true values that are based on the real data set and average empirical values across simulated data are tabulated. The results demonstrate very close agreement between specified values and empirically computed ones for all parameters under consideration. We also obtained well-known accuracy and precision quantities such as standardized bias, relative bias, average width of 95% confidence intervals, coverage rate, and root mean square error (not reported for brevity), following the evaluation system of Demirtas [21–25]; additional results are available upon request.

We briefly discuss the limitations and the advantages of the suggested algorithm and future directions in the next section.

5. Discussion

The proposed method is concerned with repeatedly generating data that on average mimic massive public health 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. A distinct advantage is that once data are generated, variables can be treated as outcomes or predictors in the subsequent analyses. Our method works well when the following occur: (i) specified

Table I. True and empirical proportions (p's) and odds ratios (OR's) for binary variables across 1000 simulation replications.

	p_S	p_M	p_G	$OR_{S,M}$	$OR_{S,G}$	$OR_{M,G}$
True value	0.5075922	0.4490239	0.5271150	1.0700000	0.9880631	0.7785208
Empirical value	0.5060613	0.4524191	0.5217678	1.0835499	0.9931258	0.7706965

S, M, and G stand for *SMOKER*, *MALE*, and *GRADE10*, respectively.

Table II. True and empirical measures of location, dispersion, symmetry, and elongation (in parenthesis) across 1000 simulation replications.

	Mean	Median	Std. dev	IQR	Skewness	Kurtosis
AVGPA	6.7774679	6.8318182	1.2355936	1.5842857	−0.24972172	0.05463359
	(6.7791157)	(6.8335664)	(1.2362081)	(1.5858131)	(−0.24522056)	(0.05676672)
AVGNA	3.4849921	3.3130435	1.5255682	2.4282883	0.48876659	−0.45673100
	(3.4853008)	(3.3136328)	1.5261456)	(2.4277109)	(0.48815078)	(−0.44770409)
NOVSEEK	2.5231020	2.5000000	0.6539071	0.8700000	−0.27853496	−0.02911116
	(2.5224734)	(2.5000378)	(0.6522825)	(0.8682042)	(−0.28383598)	(−0.03338476)
NEGMR	2.4635575	2.4300000	0.6799974	0.9300000	−0.16148671	−0.36262340
	(2.4635167)	(2.4290397)	(0.6797996)	(0.9263114)	(−0.16371828)	(−0.35905341)
ALONE	0.5167112	0.5217391	0.1964618	0.2788462	−0.02951472	−0.53345911
	(0.5165767)	(0.5215235)	(0.1958537)	(0.2764578)	(−0.03292299)	(−0.54074564)

IQR, interquartile range.

Table III. True and empirical (in parenthesis) correlations across 1000 simulation replications.

	AVGNA	*NOVSEEK*	*NEGMR*	*ALONE*	*SMOKER*	*MALE*	*GRADE10*
AVGPA	−0.5679	0.0631	0.3444	−0.2139	−0.0699	0.1126	0.0055
	(−0.5669)	(0.0629)	(0.3454)	(−0.2144)	(−0.0704)	(0.1122)	(0.0057)
AVGNA		0.0763	−0.3827	0.0893	0.0876	−0.1994	0.0264
		(0.0759)	(−0.3830)	(0.0891)	(0.0876)	(−0.1989)	(0.0259)
NOVSEEK			0.0326	−0.0439	0.0862	−0.0742	−0.0417
			(0.0325)	(−0.0437)	(0.0859)	(−0.0749)	(−0.0415)
NEGMR				0.0195	−0.0101	0.2326	0.0434
				(0.0195)	(−0.0100)	(0.2335)	(0.0432)
ALONE					0.0139	0.2217	0.0589
					(0.0136)	(0.2219)	(0.0584)
SMOKER						0.0168	−0.0032
						(0.0168)	(−0.0030)
MALE							−0.0617
							(−0.0621)

correlation matrix (Σ) is positive semidefinite; (ii) correlation bounds among variables are not violated; and (3) symmetry-peakedness (skewness-elongation) behavior for continuous variables is within the region that can be handled by power polynomials. In addition, even when these conditions hold, the final correlation matrix (Σ^*) may not be positive semidefinite. In such cases, one may resort to the nearest positive semidefinite matrix. On another note, this approach is currently designed to accommodate linear associations; extensions to model non-linear associations will be taken up in future work. Furthermore, if the real data set has missing values, one can simulate incomplete data set by first generating full data and then imposing missing values by some non-response mechanism [26–28]. Finally, ideas presented in this paper can be incorporated into the multivariate ordinal data generation algorithm proposed by Demirtas [29] to produce binary–ordinal–continuous combinations.

Considering its computational simplicity, generality, and flexibility, the suggested method is likely to be a handy addition to a 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 public health researchers need to regenerate the original data trends in simulated environments, they could implement this technique in their favorite platform and software with relative ease.

Acknowledgements

The project described was supported by award numbers P01CA098262 and R21CA140696 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. Emrich JL, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. *The American Statistician* 1991; **45**:302–304.
2. Demirtas H, Doganay B. Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics* 2012; **22**:223–236.
3. Headrick TC. *Power Method Polynomials and Other Transformations*. Chapman and Hall/CRC: Boca Raton, Florida, 2010.
4. Fleishman AI. A method for simulating non-normal distributions. *Psychometrika* 1978; **43**:521–532.
5. Vale CD, Maurelli VA. Simulating multivariate nonnormal distributions. *Psychometrika* 1983; **48**:465–471.
6. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 2003; **90**:455–463.
7. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
8. Headrick TC, Sawilowsky SS. Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics* 2000; **25**:417–436.
9. Demirtas H, Hedeker D. Multiple imputation under power polynomials. *Communications in Statistics—Simulation and Computation* 2008; **37**:1682–1695.

10. Headrick TC. Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis* 2002; **40**:685–711.
11. Burr IW. Cumulative frequency functions. *Annals of Mathematical Statistics* 1942; **13**:215–232.
12. Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika* 1949; **36**:149–176.
13. Parrish RS. Generating random deviates from multivariate Pearson distributions. *Computational Statistics and Data Analysis* 1990; **9**:283–295.
14. Schmeiser BW, Deutch SJ. A versatile four parameter family of probability distributions suitable for simulation. *AIIE Transactions* 1977; **9**:176–182.
15. Tadikamalla PR. On simulating non-normal distributions. *Psychometrika* 1980; **45**:273–279.
16. Headrick TC, Kowalchuk RK. The power method transformation: its probability density function, distribution function, and its further use for fitting data. *Journal of Statistical Computation and Simulation* 2007; **77**:229–249.
17. Demirtas H, Hedeker D. A practical way for computing approximate lower and upper correlation bounds. *The American Statistician* 2011; **65**:104–109.
18. Higham NJ. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis* 2002; **22**:329–343.
19. Bolger N, Davis A, Rafaeli E. Diary methods: capturing life as it is lived. *Annual Review of Psychology* 2003; **54**:579–616.
20. Hedeker D, Mermelstein RJ, Demirtas H. An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics* 2008; **64**:627–634.
21. Demirtas H. Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica* 2004; **58**:466–482.
22. Demirtas H. 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* 2007; **36**:871–889.
23. Demirtas H, Arguelles LM, Chung H, Hedeker D. On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics and Data Analysis* 2007; **71**:4064–4068.
24. Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation* 2008; **78**:69–84.
25. Demirtas H, Hedeker D. An imputation strategy for incomplete longitudinal ordinal data. *Statistics in Medicine* 2008; **27**:4086–4093.
26. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2003; **22**:2553–2575.
27. Demirtas H. Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2005; **24**:2345–2363.
28. Demirtas H, Hedeker D. Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine* 2007; **26**:782–799.
29. Demirtas H. A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation* 2006; **76**:1017–1025.