Published in final edited form as:

J Stat Comput Simul. 2016 ; 86(18): 3595–3607. doi:10.1080/00949655.2016.1177530.

Concurrent generation of multivariate mixed data with variables of dissimilar types

Anup Amatya^{*,a} and Hakan Demirtas^b

^aDepartment of Public Health Sciences, New Mexico State University, 1335 International mall, RM 102, Las Cruces, NM 88011

^bDivision of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 W. Taylor St., Chicago, IL, 60612

Abstract

Data sets originating from wide range of research studies are composed of multiple variables that are correlated and of dissimilar types, primarily of count, binary/ordinal and continuous attributes. The present paper builds on the previous works on multivariate data generation and develops a framework for generating multivariate mixed data with a pre-specified correlation matrix. The generated data consist of components that are marginally count, binary, ordinal and continuous, where the count and continuous variables follow the generalized Poisson and normal distributions, respectively. The use of the generalized Poisson distribution provides a flexible mechanism which allows under- and over-dispersed count variables generally encountered in practice. A step-by-step algorithm is provided and its performance is evaluated using simulated and real-data scenarios.

Keywords

Generalized Poisson; Multivariate ordinal; discretization

1. Introduction

In many research and application areas, data sets including only one type of variables is highly unusual. Most data sets often contain a combination of binary, ordinal, count, and continuous variables that are correlated with complex dependence structure. That is, mixed data sets are not uncommon in wide range of scientific fields including in health and behavioral sciences, social sciences, and economics. For example, the Interstitial Cystitis database [1] includes longitudinal data on pain in the pelvic or bladder area, urgency (pressure to urinate), urinary frequency, nocturnal void, and nocturia from a prevalent cohort of subjects with Interstitial Cystitis. For each individual, these outcomes were associated with covariates, such as the demographic and clinic characteristics of the patients, which were observed at different times. The pain and urgency scores are continuous variables, urinary frequency is a discrete count variable, nocturnal void is a three-level ordinal variable, and nocturia is a dichotomous variable. The primary aim of the study was to determine

^{*}Corresponding author. aamatya@nmsu.edu.

whether symptoms tend to co-fluctuate together indicating a single underlying aetiology or vary independently suggesting multiple mechanisms at work. Due to the ubiquitousness of mixed outcomes, development of flexible models and methods for joint analysis of such data is an active area of research [2, 3]. A Monte-Carlo simulation to assess the performance of such models and methods typically require a mechanism to generate multivariate data that closely resemble real data sets that consist of mixed variable attributes.

Multivariate data generation has been addressed extensively in statistical literature. Krummenauer [4], Cai and Kendall [5], Minhajuddin et al. [6], Shin and Pasupathy [7], Yahav and Shmueli [8], and Barbiero and Ferrari [9] have developed various methods for generating simulated data from multivariate Poisson distribution. The Barbiero and Ferrari [9] approach was shown to be comparatively more efficient, user-friendly, and often more accurate than the prior methods. Biswas [10], Demirtas [11], and Ferrari and Barbiero [12] have described and evaluated procedures for multivariate ordinal data generation. Recently, a few notable developments have occurred in the area of mixed data generation. Ruscio and Kaczetow [13] has proposed a robust iterative technique that can generate mixed data either mimicking the marginal distribution observed in the sample data or utilizing a pre-specified parameters. Demirtas and Doganay [14], Demirtas et al. [15], Demirtas and Yavuz [16] and Amatya and Demirtas [17] have advanced the field by developing and implementing methods for simulating correlated mixed data comprised of binary-normal, binary-nonnormal, ordinal-normal, and count-normal components, respectively.

This work is motivated by the need of simultaneously simulating data composed of variables with different attributes, of continuous, count and binary/ordinal types that are often encountered jointly in many settings. The proposed mechanism is built upon a combination of a few random variate generation methods that involve simulation of multivariate count data that follow the generalized Poisson distribution, multivariate binary/ordinal data, multivariate continuous data that follow normal distribution, and a mix of these distributions. The proposed method maintains the specified marginal characteristics of each variable as well as the linear association structure among them. The methods in Famoye [18], Ferrari and Barbiero [12], and Barbiero and Ferrari [9] are particularly relevant for the development of the current work. We describe the details of these methods in Sections 2–4. Furthermore, since binary is a special case of ordinal, in what follows ordinal is meant to include binary as well.

A remark on the use of the generalized Poisson distribution (GPD) in this article is in order. Generally, count data are assumed to follow a Poisson distribution which requires the mean-variance equality. However, real data in practice rarely meet such a restriction. Most often, a variance of the count data is larger than its mean (over-dispersion) and seldom smaller than the mean (under-dispersion) [19–21]. Such phenomenon is common in longitudinal count data either by marginal behavior of these variables or the correlation induced by the repeated nature of the count variables. For example, Thall and Vail [22] gives a data set on two-week seizure counts for 59 epileptics. At each of four successive post randomization clinic visits, the number of seizures occurring over the previous two weeks was reported. The mean (variance) of the number of seizures in each of the four clinical visits were 8.95 (33.92), 8.36 (26.29), 8.44 (35.69) and 7.34 (22.14). Clearly, the basic equal mean and variance

assumption of the ordinary Poisson distribution is violated. Thus, an ordinary Poisson distribution is not appropriate to model such over-or under-dispersed count data. A better choice is the generalized Poisson distribution which allows both over- or under-dispersion.

The article is organized as follows. In Section 2, we provide background information on GPD and computation of its quantile function. In Section 3, we describe a process involved in transforming a univariate normal random variable into a discrete random variable. In Section 4, we establish connection between both discrete and continuous variables by adjusting Barbiero and Ferrari [9] method for GPD and deriving an expression for correlation between discretized and normal variables. In Section 5, we outline the proposed methodology and algorithm to generate multivariate mixed data with count (GPD), ordinal, and continuous (normal) components. In Section 6, we present simulation studies for assessing the performance of the suggested method. In Section 7, we illustrate the utility of the method in mimicking the Interstitial Cystitis data. In Section 8, we conclude the paper with discussions, remarks, and future directions.

2. Generalized Poisson Distribution

The generalized Poisson distribution is a two parameter distribution that includes the ordinary Poisson distribution as a special case. Let X be a positive discrete random variable, whose probability mass function $P_x(\theta, \lambda)$ is defined as

$$P_x(\theta, \lambda) = \begin{cases} \frac{\theta(\theta + \lambda_x)^{(x-1)}}{x!} e^{-\theta - \lambda_x} & \text{for } x=0, 1, 2, \dots \\ 0 & \text{for } x > q \text{ when } \lambda < 0 \end{cases} \quad (1)$$

and zero otherwise, where $\theta > 0$, $\max(-1, -\theta/q) \leq \lambda < 1$ and $q \geq 4$ is the largest positive integer for which $\theta + \lambda q > 0$ when $\lambda < 0$. Then, X is said to follow GPD. The GPD reduces to the Poisson distribution when $\lambda = 0$ and possesses the dual characteristics of over- and under-dispersion depending on whether $\lambda > 0$ or $\lambda < 0$, respectively. The first four moments of GPD (as derived in Ch. 9 of [23]) are as follows:

$$\gamma_1 = \frac{\theta}{1 - \lambda}, \quad \gamma_2 = \frac{\theta}{(1 - \lambda)^3}, \quad \gamma_3 = \frac{1 + 2\lambda}{\sqrt{\theta(1 - \lambda)}}, \quad \gamma_4 = \frac{1 + 8\lambda + 6\lambda^2}{\theta(1 - \lambda)}. \quad (2)$$

The cumulative distribution function $F_x(\theta, \lambda)$ and quantile function $F_x^{-1}(p|\theta, \lambda)$ of GPD are defined as:

$$F_x(\theta, \lambda) = P(X \leq k|\theta, \lambda) = \sum_{x=0}^k P_x(\theta, \lambda) \quad (3)$$

$$F_x^{-1}(p|\theta, \lambda) = \min(x \in \mathbb{N}_0 : F_x(x) \geq p), \quad p \in (0, 1) \text{ and } \mathbb{N}_0 = \{0, 1, 2, \dots\}. \quad (4)$$

The expressions (3) and (4) can be evaluated efficiently by exploiting the following recurrence relation between GPD probabilities.

$$P_x(\theta, \lambda) = \frac{\theta - \lambda + \lambda x}{x} \left(1 + \frac{\lambda}{\theta - \lambda + \lambda x} \right)^{x-1} e^{-\lambda} P_{x-1}(\theta, \lambda) \text{ for } x \geq 1 \text{ and } P_0(\theta, \lambda) = e^{-\theta}. \quad (5)$$

We are not aware of readily available software implementation for $F_x^{-1}(q; \theta, \lambda)$ computation (SAS provides functions only for $\lambda > 0$). We modify Famoye [18] method—developed for generating univariate generalized Poisson random variate—and adapt it for the computation of F_x^{-1} . The modified algorithm is as follow:

- a. Set $x = 0$, $w = e^{-\lambda}$, $S = e^{-\theta}$, and $P = S$
- b. while $q > S$, do

$x = x + 1$
 $C = \theta - \lambda + \lambda x$
 $P = wC(1 + \lambda/C)^{(x-1)}P/x$
 $S = S + P$
- c. Deliver x

3. Discretization

A discrete random variable can be obtained by discretization of any continuous random variable. The simple and precise approach is to generate a univariate discrete random variable with known probability mass function through discretization of a continuous normal random variable. Let Z be a standard normal random variable. A discrete random variable $X \in \{x_1, x_2, \dots, x_{max}\}$ having distribution function F_X is constructed as follows:

$$\begin{aligned} \text{if } Z < r_1 &\rightarrow X = x_1 \\ &\vdots \\ \text{if } r_{k-1} \leq Z &\rightarrow X = x_{max}, \quad (6) \end{aligned}$$

where $r_j = \Phi^{-1}(F_X(x_j))$, Φ being the standard normal cumulative distribution function (CDF). The numerical value of x_{max} is trivial for the ordinal variable with k categories. However, for discrete count variables with a non-finite support, the largest value is not directly available. To circumvent this problem, Barbiero and Ferrari [9] suggested reducing the support of the discrete components by approximating), $x_{max} = F_x^{-1}(1 - \varepsilon)$ where ε is chosen to be as small as possible.

4. Correlation among variables with different attributes

The proposed method is concerned with simulation of data sets that are comprised of count (GPD), ordinal and normal random variables. The principle behind the proposal is to start with multivariate normal distribution and discretize its components thereby transforming them into count and ordinal components [9, 12]. In so doing, starting pairwise correlations are adjusted such that the final correlation matrix (after discretization) is close to the target correlation matrix. The details of this correlation adjustment process is given in the Section 5. Three sets of correlations are of interest: correlation among discrete random variables (GPD-GPD, GPD-ordinal, ordinal-ordinal); correlation among continuous random variables (normal-normal); and correlation among discrete and continuous random variables (GPD-normal, ordinal-normal). Of the three sets, correlations among normal random variables do not pose a major challenge as the method is built up on the concept of discretization of multivariate normal random variables.

4.1. Correlation for discrete random variables

Suppose $\mathbf{Z}$ is a multivariate normal random vector with correlation matrix Σ_Z whose discretized components form a new multidimensional discrete random variable $\mathbf{X}$ with correlation matrix Σ_X . The elements of Σ_X obviously differ from those of Σ_Z , but the transformation is tractable. Ferrari and Barbiero [12] has established the link between pairwise correlation among normal components (δ_C) and those of corresponding discretized components (δ_D) by the following expression.

$$\delta_D = \frac{\sum_{l=1}^{k_1} \sum_{t=1}^{k_2} (l - E(X_1))(t - E(X_2)) f_{Z_1 Z_2}(l/k_1, t/k_2)}{\sum_{l=1}^{k_1} (l - E(X_1))^2 (1/k_1) \sum_{t=1}^{k_2} (t - E(X_2))^2 (1/k_2)}$$

where

$$\begin{aligned} f_{Z_1 Z_2}(l/k_1, t/k_2) &= \Phi_{Z_1, Z_2}(\Phi_Z^{-1}(l/k_1) \Phi_Z^{-1}(t/k_2)) \\ &\quad - \Phi_{Z_1, Z_2}(\Phi_Z^{-1}(l/k_1) \Phi_Z^{-1}((t-1)/k_2)) - \Phi_{Z_1, Z_2}(\Phi_Z^{-1}((l-1)/k_1) \Phi_Z^{-1}(t/k_2)) \\ &\quad + \Phi_{Z_1, Z_2}(\Phi_Z^{-1}((l-1)/k_1) \Phi_Z^{-1}((t-1)/k_2)), \end{aligned} \quad (7)$$

k_1 and k_2 are the largest finite integer in the support of the discrete variables, and Φ_Z and Φ_{Z_1, Z_2} are the univariate and bivariate standard normal CDF with δ_C correlation, respectively. Ferrari and Barbiero [12] and Barbiero and Ferrari [9] used this relationship to construct correlated ordinal and correlated Poisson random variables, respectively.

4.2. Correlation Between Discrete and Normal Continuous Random Variables

Let Z_1 and Z_2 be two components of a bivariate standard normal random vector with a pairwise correlation coefficient δ_N . Suppose Z_2 is discretized to form a discrete variable X_2 , following the procedure described in the Section 3, according to a probability distribution function F_X (as defined in Equation 3). The resulting pairwise correlation δ_{ND} between Z_1 and X_2 can be calculated using a bivariate normal distribution result $E(Z_1|r_1 < Z_2 < r_2) = \delta_N E(Z_2|r_1 < Z_2 < r_2)$ and the following property of a truncated normal distribution:

$$E(Z_2|r_1 < Z_2 < r_2) = \delta_N \frac{\phi(r_1) - \phi(r_2)}{\Phi(r_2) - \Phi(r_1)}, \quad (8)$$

such that,

$$\begin{aligned} \delta_{ND} &= \frac{E(X_2)E(Z_1|X_2) - E(X_2)E(Z_1)}{\sqrt{E[X_2 - E(X_2)]^2 E[Z_1 - E(Z_1)]^2}} \\ &= \frac{\sum_{i=1}^k x_i P_X(X_2=x_i) E(Z_1|r_{i-1} < Z_2 < r_i)}{\sqrt{\sum_{i=1}^k x_i^2 P_X(X_2=x_i) - \left[\sum_{i=1}^k x_i P_X(X_2=x_i)\right]^2}} \\ &= \delta_N \frac{\sum_{i=1}^k x_i [\phi(r_{i-1}) - \phi(r_i)]}{\sigma \sqrt{\sum_{i=1}^k x_i^2 [\Phi(r_i) - \Phi(r_{i-1})] - \left[\sum_{i=1}^k x_i \{\Phi(r_i) - \Phi(r_{i-1})\}\right]^2}}, \quad (9) \end{aligned}$$

where ϕ and Φ are the standard normal probability density function (PDF) and CDF, respectively and $r_i = \Phi^{-1}(F_X(x_i))$ with $r_0 = -\infty$. For k -category ordinal variable, x_k is the last category. For a count variable following GPD, x_k is the value of the largest count as estimated by $x_k = F_x^{-1}(1 - \varepsilon)$.

In the next section, we provide a unified framework on the joint generation of count, ordinal, and normal variables. Caveats pertaining to the correlation bounds for discretized ordinal and count variables within and between variable types also apply to the proposed method. These bounds are conveniently approximated through the “generate, sort, and correlate” technique [24].

5. Algorithm for generating multivariate mixed data with generalized Poisson, ordinal, and normal components

Let $C_1, C_2, \dots, C_j$ be a set of count variables following GPD with corresponding parameter vectors $(\theta_1, \lambda_1), (\theta_2, \lambda_2), \dots, (\theta_j, \lambda_j)$; $O_1, O_2, \dots, O_m$ be a set of ordinal variables with proportion vectors $p_1, p_2, \dots, p_m$ and $W_l \sim N(\mu_l, \sigma_l^2)$, where $l = 1, 2, \dots, h$. The target linear association among three types of variables is specified in $(j + m + h) \times (j + m + h)$ correlation matrix is Σ . Without loss of generality, assume that the first j columns of data matrix consist of count variables, followed by m columns of ordinal variables, and the last h columns of normal continuous variables. Then, Σ is comprised of six components: Σ_{CC} ,

Σ_{OO} , Σ_{NN} , Σ_{OC} , Σ_{ON} and Σ_{CN} , where C , O and N correspond to count, ordinal, and normal, respectively. In this setup, each O_i may have a different number of categories or subset of them may have the same number of categories with the same or different probability vectors. Suppose that discretization of first $j + m$ components of $n \times (j + m + h)$ dimensional data matrix drawn from $N_{(j+m+h)}(\mathbf{0}, \Sigma^*)$ results into a $n \times (j + m + h)$ dimensional mixed data matrix that has the target correlation matrix Σ .

1. Compute, $x_{i_{max}} = F_{C_i}(1 - \epsilon)$ for $i = 1, \dots, j$, where F_{C_i} is as defined in the Equation 3.
2. Compute Σ^* . To this end, set up following loop to solve Equation 7 and 9 until a desired result is achieved.
 - a. Begin by setting $\kappa = 0$ and $\Sigma^*(\kappa) = \Sigma$. Let δ_N^{st*} be the st^{th} element of Σ^* for $s, t = j + m + h$.
 - b. For each elements δ_N^{st*} in $\sum_{CC}^*(\kappa)$, $\sum_{OO}^*(\kappa)$ and $\sum_{OC}^*(\kappa)$ calculate $\delta_D^{st}(\kappa)$ using Equation 7; where $s, t = 1, \dots, (j + m)$ and $s > t$.
 - c. Form a matrix $\Sigma_D(\kappa)$ by appropriately arranging $\delta_D^{st}(\kappa)$ and $\delta_{ND}^{st}(\kappa)$;
 - d. For each elements $\delta_N^{st*}(\kappa)$ in $\sum_{CN}^*(\kappa)$ calculate δ_{ND}^{ij} and $\sum_{ON}^*(\kappa)$ using Equation 9; where $s = (j + m + 1), \dots, (j + m + h)$, $t = 1, \dots, (j + m)$ and $s > t$.
 - e. Check if $|\delta^{st}(\kappa) - \delta^{st}| < \eta$ for $s = 1, \dots, (j + m + h)$, $t = 1, \dots, (j + m)$ and $s > t$, where $\delta^{st}(\kappa)$ and δ^{st} are the st^{th} element of $\Sigma_D(\kappa)$ and Σ , respectively; and η is a small positive number (say 0.00001)

If yes, set $\Sigma^* = \Sigma^*(\kappa)$. Go to ((3))

If not, calculate new δ_N^{st*} as follows

$$\delta_N^{st*}(\kappa+1) = \delta_N^{st*}(\kappa) \frac{\delta^{st}}{\delta^{st*}(\kappa)},$$

and update $\Sigma^*(\kappa)$ to form $\Sigma^*(\kappa + 1)$ such that

$\delta_N^{st*}(\kappa+1)$ constitute its elements. Set $\kappa = \kappa + 1$ and repeat ((2)(b) to ((2)(e))

3. Draw n samples from $N_{(j+m+h)}(\mathbf{0}, \Sigma^*)$. These samples form $n \times (j + m + h)$ intermediate data matrix Y^* .
4. Discretize the first j th columns of Y^* according to $F_{C_i}(\theta_i, \lambda_i)$ and $x_{i_{max}}$.

5. Discretize the $(j+1)^{th}$ to $(j+m)^{th}$ columns of $\mathbf{Y}^*$ according to proportion vectors $p_1, p_2, \dots, p_m$.
6. Center and scale the last n columns of $\mathbf{Y}^*$ according to the target mean and standard deviation vectors.

Some numerical issues may arise when implementing the algorithm. First, the target pairwise correlation between two discrete variables and correlation between discrete and continuous variables must respect the correlation bounds. These bounds can be calculated using “generate, sort, and correlate” technique [24]. Such violations need to be checked and addressed prior to Σ^* computation. Secondly, the computed intermediate correlation matrix Σ^* may not be positive definite. In such situation, one can compute the nearest positive definite matrix by the methods proposed by Higham [25, 26] using *nearPD* function in the Matrix package in R. Finally, for some parametric combinations of GPD, the F_X calculation may require comparatively longer computation time.

6. Method evaluation via simulation

The proposed method involves simulation of correlated multivariate count, ordinal, and normal variables. It requires specification of parameters governing the marginal distributions of each type of variables and also the linear association among all the variables in the form of correlation matrix. For the purpose of this evaluation we consider data sets that are composed of two GPD variables, two ordinal variables and two normal variables, i.e.; total of 6 variables. The following parameter vectors are considered for GPD variables: $\theta \in \{(1, 2), (10, 20)\}$ and $\lambda \in \{(.1, .2), (-.1, -.2), (.4, .6), (-.2, .4), (-.2, -.4), (.5, -.7)\}$. These parameter values represent moderate to high over- and under-dispersed count variables. To avoid a proliferation of simulation scenarios, we have limited this simulation for a single set of parameter values of ordinal and normal components. The cumulative marginal probabilities of two ordinal components—one with three categories and another with four categories—were chosen to be $p_1 = (.35, .80)$ and $p_2 = (.32, .54, .81)$. This specification covers a reasonably broad range of probability values. The means and standard deviations of the normal components are set at $(0, 1)$. Finally, the correlation matrices are specified randomly. The values of pairwise correlations are checked for bound violation, symmetry, and positive definiteness, ensuring the validity of the randomly generated correlation matrix. Performance of the proposed method is evaluated for each combination of parameters using the criteria described in the next section.

6.1. Accuracy and precision

We report the relative bias (RB) of the estimates $\left(\frac{E(\hat{\eta}) - \eta}{\eta} \times 100\%\right)$ over 1000 Monte Carlo replications, where η is the true value of a parameter and $\hat{\eta}$ is the corresponding estimated value. Also, the root mean square error RMSE ($\hat{\eta}$) is estimated using $\sqrt{E(\hat{\eta} - \eta)^2}$, the standardized bias (SB) is estimated using, $\left(\frac{|\sum(\hat{\eta}) - \eta|}{SD(\hat{\eta})} \times 100\%\right)$ and coverage rate (CR) is defined as a percentage of times that η is contained within a 95% confidence interval. Finally, RB and SB are the measures of accuracy and RMSE and CR are the

integrated measures of some combination of accuracy and precision. Typically, we desire $RB < 5\%$ and $SB < 50\%$ [27].

6.2. Results

We generate 1000 sets of simulated data ($n=200$) for each combinations of parameters. A subset of results of the evaluation is shown in Table 1 due to space limitation. The full table is available at the second author's website, <http://demirtas.people.uic.edu/table1.pdf>. The mean pairwise correlations are strikingly close to the target values for all combination of parameter specifications. The values on all the evaluation criteria are well within the tolerable limits. The average cumulative proportions of two ordinal variables are also very close to the specified target values ($RB < 1\%$, $SB < 8\%$, $CR \approx 95\%$). Furthermore, $RB < 1.4\%$ and $SB < 11\%$ for count variables indicate the accurate replication of the specified values of mean (γ_1) and standard deviation ($\sqrt{\gamma_2}$) of GPD; and $CR \approx 95\%$ suggest high precision. The most difficult parameters to recover accurately are the higher moments—skewness (γ_3) and kurtosis (γ_4) of the GPD variables, especially for low θ and high λ combinations. For skewness parameters, $RB < 9\%$ and $SB < 35\%$ indicating acceptable results with somewhat low accuracy, whereas $CR \approx 95\%$ suggest acceptable precision. Similarly, for kurtosis parameters, $RB < 19\%$, $SB < 42\%$, and $CR 94\%–97\%$. Given that only the fourth moment of GPD variable, under extreme conditions (small θ and large λ), produces values on evaluation metrics that are over thresholds—in our opinion—the proposed method does a remarkable job at generating mixed data with count, ordinal, and normal components, as measured by the concordance between the specified and empirically computed quantities on average.

We do not report the results of small sample cases for brevity, as expected mean, RB, SB did not change substantially. The CR was somewhat higher, but not to a degree of inefficiency. In addition, as one would expect, RMSE was also slightly larger.

7. Example: Interstitial Cystitis Data Base study

We consider the Interstitial Cystitis Data Base (ICDB) study to illustrate the performance of the proposed method to mimic real data. We introduced this data in the Introduction section. Briefly, the ICDB was established by National Institute of Diabetes and Digestive and Kidney Disease in 1993 to understand the natural history of the interstitial cystitis (IC) and to assess the demographic and clinical characteristics of IC patients [28]. A number of articles have been published utilizing this database since its establishment. Among other variables, the data set contains information on number of non-nocturnal voids per day (VF_AWAKE), ordinal levels of urinary urgency over last four weeks (PURG2C3), and average maximum interval between voids per day (AVGMINT). For this illustration, we selected only the patients who had complete data on all three variable across 3-, 6- and 9-month visits. A total of 309 patients met this requirement. The final data excluded variables that are not under consideration and patients that did not meet the aforementioned requirement. The VF_AWAKE variable measured on three occasions constitute three correlated count variables; PURG2C3 variable measured on three occasions constitute three correlated ordinal (three levels) variables; and AVGMINT also measured on three occasions

constitute three correlated normal variables. Together, they form a data set with nine correlated variables of dissimilar types.

We demonstrate an ability of the proposed method to duplicate this “model” data in two ways. First, we show the proximity of the four moments of VF_AWAKE, cumulative proportions of PURG2C3, mean and standard deviation of AVGMINT, and the correlation matrix estimated from the “model” data and from simulated data (mimicking the “model” data). From the “model” data, θ and λ parameters for VF_AWAKE are obtained by fitting univariate generalized Poisson distribution to 3-, 6- and 9-month assessment data. The appropriateness of marginal normal distributions for AVGMINT variables are tested using the Anderson-Darling normality test ($p\text{-value} > 0.05$). The correlation matrix and the values of the relevant parameters estimated from the “model” data are presented on the Tables 2 and 3, respectively. Using these estimated quantities as true parameters, the process of generating a data matrix of size 309×9 is replicated 1000 times.

The results of the comparison of these quantities between the “model” and simulated data are presented on the Figures 1 – 4. Figure 1 shows three evaluation metrics. The maximum observed values (representing the worst-case scenario) on all three metrics fall well under the thresholds for all the parameters including kurtoses of count variables. Figure 2 shows distributions of cumulative proportion of each categories of ordinal variables PURG2C3 assessed over 3 time points. The x-axis displays the “true” values from the “model” data and the y-axis displays estimated values from simulated data sets. The average estimated cumulative proportions are in close proximity to the “true” values with relatively narrow interquartile range. Figure 3 shows 36 estimated mean pairwise correlations along with their “true” value on the x-axis. It shows that the average estimated pairwise correlations over simulated data are in close vicinity of the “true” values. The largest difference between the average and “true” pairwise correlations is less than 0.0022. Finally, Figure 4 presents the distributions of four sample moments of VF_AWAKE variables. The upper two graphs show the distributions of sample means and standard deviations. The averages of these two statistics are close to the “true” values with very few outliers. As expected, distributions of the third and the fourth moments contain more outliers. However, the averages are fairly close to their “true” values. Overall, the proposed method succeed remarkably in duplicating the marginal distributions and dependence structure of the “model” data.

Next, we fit a mixed-effect ordinal regression model with cumulative probit link and equally spaced thresholds. We show that the estimated values of the parameters of this regression model using the “model” data and simulated data are close to each other. The model under consideration is as follows:

$$\phi^{-1}\{P(y_{ij} < c)\} = \alpha_c + \beta_0 AVGMINT + \beta_1 VF_AWAKE + \beta_2 MONTH + v_{0i}, \quad (10)$$

where y_{ij} is an unobservable latent variable related to the ordinal variable PURG2C3, α_c is a series of threshold values, and v_{0i} is a random effect assumed to have $N(0, \sigma_{v0})$ distribution. Table 4 presents the estimated regression parameters for the “model” data on the “True Value” column and for the simulated data on “Mean” column along with

evaluation metrics on the remaining columns. Clearly, the average of the parameter estimates from 1000 simulated data sets are well within the tolerable limits. These results suggest that the proposed method is useful in capturing and reconstructing the real data trends, and may be regarded as a useful tool for evaluating statistical models and methods.

8. Discussion

The approach to multivariate data generation with mixed variable types in this paper is motivated by the emerging popularity of data analysis techniques for mixed outcomes. The method draws ideas from previous works by Ferrari and Barbiero [12], Barbiero and Ferrari [9], Demirtas and Doganay [14], and Amatya and Demirtas [17]. The novelties in this paper are the use of the generalized Poisson distribution to address over- and under-dispersion in count data, derivation of a connection (correlation) between normal and count/ordinal variables obtained via discretization, and the unified framework through which data sets composed of dissimilar variable types are generated. The proposed framework is also capable of generating multivariate data composed of only one type of variables mentioned in this work. This is particularly useful for generating multivariate GPD variables, method for which is not currently available. For the values of parameters that are generally encountered in practice, the proposed method performs considerably well and provides a robust mechanism for generating multivariate data with combination of count, ordinal and normal variables. Future work will address modeling continuous variable components in the mixed data via non-normal continuous distributions.

References

1. Simon LJ, Landis JR, Erickson DR, Nyberg LM, Group IS, et al. The interstitial cystitis data base study: concepts and preliminary baseline descriptive statistics. *Urology*. 1997; 49(5):64–75. [PubMed: 9146004]
2. Song P, Li M, Yuan Y. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*. 2009; 65(1):60–68. [PubMed: 18510653]
3. De Leon, AR.; Chough, KC. Analysis of mixed data: Methods and applications. CRC Press; 2013.
4. Krummenauer F. Efficient simulation of multivariate binomial and Poisson distributions. *Biometrical Journal*. 1998; 40(7):823–832.
5. Cai Y, Kendall W. Perfect simulation for correlated Poisson random variables conditioned to be positive. *Statistics and Computing*. 2002; 12(3):229–243.
6. Minhajuddin TM, Harris IR, Schucany WR. Simulating multivariate distributions with specific correlations. *Journal of Statistical Computation and Simulation*. 2004; 74(8):599–607.
7. Shin, K.; Pasupathy, R. Proceedings of the 39th Conference on Winter Simulation: 40 Years! The Best is Yet to Come; WSC '07. Piscataway, NJ, USA: IEEE Press; 2007. A method for fast generation of bivariate Poisson random vectors; p. 472-479.
8. Yahav I, Shmueli G. On generating multivariate Poisson data in management science applications. *Applied Stochastic Models in Business and Industry*. 2012; 28(1):91–102.
9. Barbiero A, Ferrari PA. Simulation of correlated Poisson variables. *Applied Stochastic Models in Business and Industry*. 2014; 31(5):669–680.
10. Biswas A. Generating correlated ordinal categorical random samples. *Statistics and Probability Letters*. 2004; 70(1):25–235.
11. Demirtas H. A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*. 2006; 76(11):1017–1025.
12. Ferrari PA, Barbiero A. Simulating ordinal data. *Multivariate Behavioral Research*. 2012; 47(4): 566–589. [PubMed: 26777670]

13. Ruscio J, Kaczetow W. Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*. 2008; 43(3):355–381. [PubMed: 26741201]
14. Demirtas H, Doganay B. Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*. 2012; 22(3):323–236.
15. Demirtas H, Hedeker D, Mermelstein JM. Simulation of massive public health data by power polynomials. *Statistics in Medicine*. 2012; 31(27):3337–3346. [PubMed: 22532052]
16. Demirtas H, Yavuz Y. Concurrent generation of ordinal and normal data. *Journal of Biopharmaceutical Statistics*. 2015; 25(4):635–650. [PubMed: 24906138]
17. Amatya A, Demirtas H. Simultaneous generation of multivariate data with Poisson and normal marginals. *Journal of Statistical Computation and Simulation*. 2015; 85(15):3129–3139.
18. Famoye F. Generalized Poisson random variate generation. *American Journal of Mathematical and Management Sciences*. 1997; 17(3–4):219–237.
19. Consul, PC.; Famoye, F. Lagrangian probability distributions. Springer; 2006.
20. Sáez-Castillo AJ, Conde-Sánchez A. A hyper-Poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*. 2013; 61:148–157.
21. Lynch HJ, Thorson JT, Shelton AO. Dealing with under- and over-dispersed count data in life history, spatial, and community ecology. *Ecology*. 2014; 95(11):3173–3180.
22. Thall PF, Vail SC. Some covariance models for longitudinal count data with overdispersion. *Biometrics*. 1990; 46(3):657–671. [PubMed: 2242408]
23. Consul, PC. Generalized Poisson distributions: Properties and applications. New York: Marcel Dekker, Inc.; 1989.
24. Demirtas H, Hedeker D. A practical way for computing approximate lower and upper correlation bounds. *American Statistician*. 2011; 65(2):104–109.
25. Higham NJ. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*. 1988; 103:103–118.
26. Higham NJ. Computing the nearest correlation matrix-a problem from finance. *IMA Journal of Numerical Analysis*. 2002; 22(3):329–343.
27. Demirtas H. Simulation driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*. 2004; 58(4):466–482.
28. Probert KJ, Schaeffer AJ, Brensinger CM, Kusek JW, Nyberg LM, Landis JR. A prospective study of interstitial cystitis: results of longitudinal followup of the interstitial cystitis data base cohort. The interstitial cystitis data base study group. *The Journal of Urology*. 2000; 163(5):1434–1439. [PubMed: 10751852]

Figure 1. Accuracy of the estimated parameters. RB: Relative bias, SB: Standardized Bias.

Figure 2.
Cumulative probabilities of three ordinal variables.

Figure 3. Correlations among variables with different attributes, of count, categorical and continuous types.

Figure 4.
Four moments of GPD count variables.

Table 1

Key descriptive statistics and evaluation quantities. The mean represents the average of estimates across 1000 simulation replicates.

θ	λ	True Value	Evaluation Criteria									
			Mean	RB	SB	RMSE	CR	Min.	1st Qu.	3rd Qu.	Max.	
(1, 2)	ρ	0.1865	0.1864	0.0707	0.1814	0.0727	96.10	-0.0383	0.1348	0.2366	0.4494	
		0.1241	0.1242	0.1104	0.1962	0.0698	95.10	-0.1069	0.0765	0.1695	0.3350	
		0.1965	0.1919	2.3207	6.6481	0.0687	95.40	-0.0762	0.1460	0.2381	0.4173	
		0.3627	0.3611	0.4283	2.5224	0.0616	95.60	0.1457	0.3182	0.4043	0.5231	
		0.2671	0.2678	0.2726	1.0978	0.0663	95.00	0.0375	0.2234	0.3129	0.4515	
		0.2783	0.2762	0.7674	3.3110	0.0645	95.50	0.0646	0.2354	0.3207	0.4802	
		0.1932	0.1892	2.0507	5.7763	0.0687	95.40	-0.0598	0.1447	0.2362	0.3784	
		0.3764	0.3733	0.8107	5.1045	0.0598	95.70	0.1704	0.3327	0.4148	0.5409	
		0.2704	0.2718	0.5161	2.1930	0.0636	96.00	0.0791	0.2290	0.3168	0.4293	
		0.2735	0.2687	1.7604	7.3842	0.0654	94.30	0.0629	0.2277	0.3131	0.5225	
		0.1174	0.1206	2.7414	4.4714	0.0720	95.60	-0.1149	0.0767	0.1684	0.3594	
		0.1438	0.1437	0.0478	0.0985	0.0697	95.00	-0.1084	0.0971	0.1915	0.3769	
		0.3861	0.3851	0.2592	1.6800	0.0596	95.20	0.1573	0.3466	0.4267	0.5876	
		0.3356	0.3332	0.7189	3.8945	0.0620	95.30	0.1255	0.2921	0.3736	0.5146	
		0.3825	0.3805	0.5100	3.1678	0.0616	95.30	0.1558	0.3404	0.4223	0.5569	
		1.1111	1.1081	0.2751	3.8193	0.0800	95.20	0.8600	1.0600	1.1600	1.3800	
		2.5000	2.5003	0.0100	0.1893	0.1320	94.60	2.1150	2.4100	2.5850	2.9900	
		1.1712	1.1675	0.3216	4.7146	0.0799	95.60	0.8931	1.1120	1.2240	1.4580	
		1.9764	1.9728	0.1827	2.7420	0.1317	94.60	1.6180	1.8780	2.0590	2.4630	
		1.2649	1.2150	3.9429	16.9577	0.2982	96.70	0.6300	1.0150	1.3670	3.8590	
		1.1068	1.0685	3.4581	13.1912	0.2925	96.30	0.4125	0.8793	1.2200	3.6660	
		5.0667	4.7724	5.8079	15.7250	1.8934	97.10	2.4800	3.7110	5.3270	31.7100	
		4.7750	4.5638	4.4222	12.3918	1.7162	96.60	2.3540	3.5790	5.0690	29.9600	
		p	0.3500	0.3517	0.4857	5.0079	0.0339	0.95	0.2450	0.3300	0.3750	0.4600

θ	λ	Evaluation Criteria									
		True Value	Mean	RB	SB	RMSE	CR	Min.	1st Qu.	3rd Qu.	Max.
ρ	(10,20) (-2,-.4)	0.8000	0.8009	0.1175	3.3734	0.0279	0.94	0.7250	0.7800	0.8200	0.8900
		0.3200	0.3206	0.1891	1.8271	0.0331	0.95	0.2100	0.3000	0.3412	0.4300
		0.5400	0.5405	0.0870	1.3241	0.0355	0.95	0.4450	0.5150	0.5650	0.6450
		0.8100	0.8115	0.1895	5.5632	0.0276	0.95	0.7150	0.7950	0.8300	0.9000
		0.3317	0.3350	1.0009	5.1817	0.0641	95.60	0.0957	0.2912	0.3784	0.6035
		0.2449	0.2464	0.6458	2.2962	0.0688	95.90	0.0429	0.1973	0.2950	0.4437
		0.2780	0.2801	0.7529	3.2329	0.0647	95.40	0.0861	0.2358	0.3262	0.5135
		0.1837	0.1846	0.4911	1.3465	0.0670	95.60	-0.0017	0.1385	0.2282	0.3998
		0.2717	0.2748	1.1611	4.8461	0.0651	95.30	0.0560	0.2318	0.3186	0.4873
		0.1865	0.1879	0.7666	2.1196	0.0674	94.50	-0.0125	0.1404	0.2332	0.3891
		0.2107	0.2122	0.7011	2.2212	0.0665	95.60	-0.0142	0.1654	0.2583	0.4039
		0.3617	0.3646	0.7905	4.6359	0.0617	95.30	0.0705	0.3226	0.4057	0.5569
		0.2135	0.2137	0.1007	0.3102	0.0693	95.90	-0.1018	0.1700	0.2608	0.4171
		0.1548	0.1540	0.5020	1.1363	0.0684	95.10	-0.0769	0.1108	0.1996	0.3751
		0.3188	0.3201	0.3926	1.9556	0.0640	95.80	0.1381	0.2768	0.3670	0.5221
		0.3587	0.3579	0.2130	1.1931	0.0640	95.20	0.1297	0.3142	0.4016	0.5430
γ^1	$(\sqrt{\gamma_2})$	0.2669	0.2645	0.8860	3.5797	0.0661	95.00	0.0444	0.2218	0.3089	0.4261
		0.2335	0.2331	0.1932	0.6729	0.0670	95.40	0.0332	0.1877	0.2808	0.4634
		0.2407	0.2426	0.7809	2.9008	0.0648	95.30	0.0498	0.1990	0.2878	0.4488
		0.1796	0.1844	2.6872	7.2200	0.0670	94.30	-0.0184	0.1388	0.2281	0.4188
γ^3		0.3926	0.3935	0.2266	1.4939	0.0595	95.00	0.1644	0.3538	0.4330	0.5537
		8.3333	8.3287	0.0557	2.7920	0.1663	94.90	7.8450	8.2200	8.4450	8.8700
		14.2857	14.2825	0.0226	1.6308	0.1980	95.00	13.6900	14.1500	14.4200	14.9000
		2.4056	2.4053	0.0136	0.2764	0.1185	95.50	2.0670	2.3200	2.4880	2.8090
γ^3		2.6998	2.6938	0.2214	4.4824	0.1334	95.60	2.3110	2.5990	2.7920	3.0490
		0.1732	0.1670	3.5937	3.8583	0.1614	95.00	-0.3318	0.0609	0.2699	0.8250
		0.0378	0.0344	9.0104	2.0669	0.1647	95.10	-0.4314	-0.0853	0.1426	0.6621

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	True Value	Evaluation Criteria								λ	θ
		Mean	RB	SB	RMSE	CR	Min.	1st Qu.	3rd Qu.		
γ^4	2.9700	2.9260	1.4826	13.6121	0.3263	96.50	2.1920	2.7030	3.0970	5.1520	
	2.9557	2.9333	0.7585	7.1579	0.3138	95.60	2.0900	2.7100	3.1080	4.1460	
	0.3500	0.3525	0.7200	7.5922	0.0332	95.60	0.2450	0.3300	0.3750	0.4500	
	0.8000	0.8009	0.1081	3.0480	0.0284	95.90	0.7150	0.7800	0.8200	0.8850	
p	0.3200	0.3203	0.1000	0.9819	0.0326	94.60	0.2150	0.3000	0.3400	0.4350	
	0.5400	0.5404	0.0722	1.0688	0.0365	95.00	0.4250	0.5150	0.5650	0.6550	
	0.8100	0.8101	0.0099	0.2760	0.0290	94.70	0.7050	0.7900	0.8300	0.8900	

Table 2

Correlation matrix of the interstitial cystitis data

	VF_AWAKE			PURG1C3			AVGMINT		
	3	6	9	3	6	9	3	6	9
months									
3	1	0.757	0.758	0.275	0.195	0.285	-0.475	-0.479	-0.466
6	0.757	1	0.806	0.246	0.274	0.312	-0.5	-0.53	-0.506
9	0.758	0.806	1	0.274	0.256	0.397	-0.466	-0.481	-0.519
3	0.275	0.246	0.274	1	0.575	0.545	-0.242	-0.241	-0.166
6	0.195	0.274	0.256	0.575	1	0.599	-0.255	-0.273	-0.201
9	0.285	0.312	0.397	0.545	0.599	1	-0.235	-0.248	-0.229
3	-0.475	-0.5	-0.466	-0.242	-0.255	-0.235	1	0.716	0.729
6	-0.479	-0.53	-0.481	-0.241	-0.273	-0.248	0.716	1	0.75
9	-0.466	-0.506	-0.519	-0.166	-0.201	-0.229	0.729	0.75	1

Table 3

Parameter estimates from interstitial cystitis data

		Assesment occations		
		3 months	6 months	9months
VF_AWAKE	θ	7.31	7.16	7.03
	λ	0.34	0.34	0.38
	γ_1	11.03	10.85	11.28
	$(\sqrt{\gamma_2})$	16.64	16.44	18.09
	γ_3	0.76	0.77	0.84
	γ_4	3.90	3.94	4.11
PURG2C3	p_1	0.37	0.40	0.41
	p_2	0.78	0.81	0.81
AVGMINT	mean	4.80	4.86	4.89
	SD	1.64	1.64	1.72

Note: SD denotes standard deviation; p_1 and p_2 denote cumulative probabilities of first and second category of three-level ordinal variable; and γ_1 , γ_2 , γ_3 and γ_4 denotes the mean, SD, skewness, and kurtosis of GPD.

Table 4

Parameter estimates of cumulative logit ordinal mixed-effect regression.

	True Value	Mean	RB	SB	RMSE	CR	Min.	1st Qu.	3rd Qu.	Max.
α_1	-0.3928	-0.3758	0.0170	4.3292	0.3495	0.95	-1.6470	-0.6098	-0.1410	0.6547
$\alpha - \alpha$	2.0762	2.0891	0.0130	0.6257	0.1204	0.94	1.7740	2.0080	2.1580	2.5180
AVGMINT	-0.2072	-0.2099	0.0027	1.2908	0.0476	0.95	-0.3864	-0.2426	-0.1782	-0.0500
VF_AWAKE	0.1186	0.1227	0.0041	3.4609	0.0165	0.95	0.0785	0.1116	0.1333	0.1799
MONTH	-0.0305	-0.0332	0.0027	8.9748	0.0186	0.94	-0.0894	-0.0457	-0.0211	0.0202
σ_{ϵ_0}	1.6228	1.5615	0.0613	3.7775	0.2670	0.95	0.8955	1.3700	1.7130	2.6300