

Flexible random effects copula models for clustered mixed bivariate outcomes

B. WU and A. R. de LEON*

*Department of Mathematics and Statistics
University of Calgary
Calgary, Alberta T2N1N4, Canada*

SUMMARY

This paper is concerned with the analysis of clustered data with mixed bivariate responses, i.e., where each member of the cluster has discrete and continuous outcomes. A copula-based random effects model is proposed that accounts for associations between discrete and/or continuous outcomes within clusters, including the intrinsic association between the mixed outcomes for the same subject. The approach yields regression parameters in models for both outcomes that are marginally meaningful; in addition, by assuming a latent variable framework to describe discrete outcomes, complications that arise from direct applications of copulas to discrete variables are avoided. The marginal distributions are flexible since we can choose any distributions for mixed outcomes. Maximum likelihood estimation of our model parameters is implemented using standard software such as PROC NLMIXED in SAS. Results of simulations concerning the bias and efficiency of the estimates are reported. The proposed methodology is motivated by and illustrated using a developmental toxicity study of ethylene glycol (EG) in mice.

Keywords: Correlated robit-normal model; Gaussian copula; Joint analysis; Likelihood estimation; Mixed binary-continuous data.

1. Introduction

Many studies in medicine and health often give rise to mixed discrete and continuous outcomes measured on the same subjects over time in the case of longitudinal studies, or from clustered subjects in cross-sectional settings. Such outcomes are typically assumed to be correlated and accounting for the correlation in the data is a common issue in the analysis. Consider, for example, the ethylene glycol (EG) data from a developmental toxicity study conducted by the National Toxicology Program (NTP) (Price et al., 1985). In the data, a total 94 pregnant mice, called

*email: adeleon@math.ucalgary.ca

dams, were randomly exposed to EG at four different dose levels, 0, 0.75, 1.5, and 3g/kg/day, during the period of development of major fetal organs. The 1028 live fetuses from the 94 litters, with litter sizes ranging from 1 to 16, were examined for various defects, including fetal weight (continuous) and the presence/absence of fetal malformations (binary). In this study, the basic sampling unit is a litter of fetuses, each yielding mixed outcomes, and the main interest is on the relationship between EG dose and the mixed outcomes. The goal is to characterize the nature of the relationship between dose and outcomes within each litter, and the association between measurements on different and/or the same fetuses.

A flexible joint model that represents these relationships – including the outcomes’ marginal and conditional distributions, and their associations – is an appropriate framework in which these questions may be addressed. However, the presence of mixed outcomes complicates the situation and conventional approaches do not directly apply in such settings. Most research dealing with statistical problems with joint analysis of mixed outcomes has appeared only recently due to a relative lack of standard models and the associated difficulties of constructing them. In statistical and practical respects, such models have many potential advantages. For example, a joint model enables analysts to account for relationships between outcomes, assess the joint influence of predictors/covariates on them, and characterize various associations at the same time. From a statistical standpoint, joint analysis avoids multiple testing and naturally leads to global tests, thus resulting in increased power and better control of Type I error rates (de Leon and Zhu, 2008). Significant efficiency gains over separate univariate analyses have also been reported, specially in settings where there are missing data (Gueorguieva and Sanacora, 2006; Fitzmaurice and Laird, 1997).

A number of approaches to joint model specification for clustered mixed outcomes have been proposed and studied in the literature (Teixeira-Pinto and Normand, 2010). Among the most popular are so-called factorization models, which are directly specified as the product of the marginal distribution of one outcome type and the conditional distribution of the other. Examples of mixed-outcome joint distributions constructed this way are general location (GLOMs) and conditional grouped continuous models (CGCMs). GLOMs (Schafer, 1997), which assume a marginal multinomial distribution for discrete outcomes and a conditional normal distribution for continuous outcomes, given discrete outcomes, have received much attention in the literature (Fitzmaurice and Laird, 1997, 1995); a recent application in oncology trials is given by Hirakawa (2011). By reversing the factorization, Catalano and Ryan (1992) proposed a CGCM-type joint model based on a normal latent description of discrete outcomes; molenberghs et al. (2001) introduced a related model based on the Plackett-Dale approach. Various refinements and extensions of these models have been studied by de Leon and Carrière (2007), George et al. (2007), Faes et al. (2004), Regan and Catalano (1999), and Catalano (1997), among others. Recent surveys are given in de Leon and Carrière (2010), Faes et al. (2009b), and Regan and Catalano (2002).

Several authors have recently adopted copulas to indirectly construct mixed-outcome joint models. Copulas, which are common in actuarial and financial applications, have proved useful in practice when the joint distribution of interest is either not available or difficult to specify but marginal distributions can be specified with confidence, like in mixed-outcome settings. Recent important references include de Leon and Wu (2011), Song et al. (2009), and Zimmer and Trivedi (2006).

Generalized linear mixed models (GLMMs) are a natural framework within which to analyze mixed-outcome data. Unlike factorization models which focus on marginal modelling, GLMMs incorporate subject-specific effects in the analysis (McCulloch, 2007). The inclusion of random effects is used to build joint models that embed an association structure between clustered measurements of a particular outcome or of different outcomes. Gueorguieva and Agresti (2001) introduced a correlated probit model for the EG data constructed from a bivariate GLMM with a probit link for the binary response and an identity link for the normally distributed continuous outcome; a recent application of the model is discussed in Najita et al. (2009) and Faes et al.

(2009a). A different but closely related model is studied in Lin et al. (2010), who incorporated correlated random effects in Fitzmaurice and Laird’s (1997, 1995) GLOM-type joint model to build in correlations between clustered binary and continuous outcomes; for interpretational ease, a bridge-distributed random effect is used for the binary outcome, and a Gaussian copula couples it with a normal random effect for the continuous outcome.

Our goal in this paper is to develop a flexible GLMM that permits non-normally distributed residual errors and random effects. The joint model is similar to Gueorguieva and Agresti’s (2001) approach in that a latent variable is adopted as well to describe the discrete outcome; however, unlike their correlated probit model, our proposal does not require either errors or random effects to be normally distributed. This is accomplished by adopting the Gaussian copula (Song, 2007) to model the joint distribution of outcomes as well as the distribution of random effects. The model is general enough to include Gueorguieva and Agresti’s (2001) and Lin et al.’s (2010) models, among others, as special cases. We adopt a fully parametric specification and use likelihood-based methods for model estimation, affording us a whole battery of available procedures for model inference, model checking, and validation. Standard statistical software/packages (e.g., PROC NL MIXED in SAS) are employed to implement likelihood estimation for the model.

This paper is organized as follows. We introduce the copula-based random effects model in Section 2. The various associations in the data (between different/same outcomes) are provided as well. In Section 3, we provide the likelihood representation of the model and discuss likelihood estimation for the model. Simulation results on the finite-sample properties of estimates are reported in Section 4. Section 5 illustrates the application of the model to the EG mice data. Finally, the paper concludes in Section 6.

2. Copula-based random effects model for mixed outcomes

Consider correlated discrete and continuous outcomes X_{ij} (e.g., presence/absence of fetal malformations) and Y_{ij} (e.g., fetal weight) for subject $j = 1, \dots, n_i$, in cluster $i = 1, \dots, N$. Suppose X_{ij} have $D + 1$ distinct values, say, $s_0 < s_1 < \dots < s_D$, possibly representing ordinal or nominal states. Let Y_{ij}^* be the unobserved continuous latent variable underlying X_{ij} , such that

$$X_{ij} = \begin{cases} s_0 & , \text{ if } Y_{ij}^* \in (-\infty, \gamma_1) \\ \vdots & \vdots \\ s_d & , \text{ if } Y_{ij}^* \in [\gamma_d, \gamma_{d+1}) \\ \vdots & \vdots \\ s_D & , \text{ if } Y_{ij}^* \in [\gamma_D, +\infty) \end{cases} \quad (1)$$

where $\gamma_1 < \dots < \gamma_D$ are unknown thresholds, with $\gamma_0 = -\infty$ and $\gamma_{D+1} = +\infty$. To simplify the ensuing discussion, we assume a shared cluster-specific random effect B_i and construct a joint probability density function (PDF) for X_{ij} and Y_{ij} as

$$f_{X_{ij}, Y_{ij}}(s_d, y_{ij}) = \int_{-\infty}^{+\infty} f_{X_{ij}, Y_{ij} | B_i}(s_d, y_{ij} | b) f_{B_i}(b) db, \quad (2)$$

for $d = 0, \dots, D$, where

$$f_{X_{ij}, Y_{ij} | B_i}(s_d, y_{ij} | b) = \int_{\gamma_d}^{\gamma_{d+1}} f_{Y_{ij}^*, Y_{ij} | B_i}(y^*, y_{ij} | b) dy^*. \quad (3)$$

In (2)–(3), $f_{Y_{ij}, Y_{ij}^* | B_i}$ is the conditional PDF of Y_{ij} and Y_{ij}^* , given B_i , and f_{B_i} is the PDF of B_i . Note that model (3) above does not require the assumption of conditional independence of the mixed outcomes, thus accommodating a direct within-subject association between them. Following de Leon and Wu (2011), we adopt the Gaussian copula to construct the conditional joint PDF $f_{Y_{ij}, Y_{ij}^* | B_i}$ in (2). In practice, statisticians oftentimes know very little about the joint behavior of outcomes but can specify their marginal behaviors reasonably well; copulas provide a useful means of assembling a joint distribution in this case. We have the conditional cumulative distribution function (CDF) of Y_{ij}^* and Y_{ij}

$$F_{Y_{ij}^*, Y_{ij} | B_i}(y_{ij}^*, y_{ij} | b) = \Phi_2(\Phi^{-1}(u_{ij}^*), \Phi^{-1}(u_{ij}); \rho_Z), \quad (4)$$

where $u_{ij}^* = F_{Y_{ij}^* | B_i}(y_{ij}^* | b)$ and $u_{ij} = F_{Y_{ij} | B_i}(y_{ij} | b)$, with $F_{Y_{ij}^* | B_i}$ and $F_{Y_{ij} | B_i}$ the conditional CDFs of Y_{ij}^* and Y_{ij} . In (4), Φ_2 is the bivariate standard normal CDF, Φ^{-1} is the standard normal quantile function, and $\rho_Z = \text{corr}(Q_{ij}^*, Q_{ij} | B_i)$ is the conditional correlation between normal scores $Q_{ij}^* = \Phi^{-1}(U_{ij}^*)$ and $Q_{ij} = \Phi^{-1}(U_{ij})$, where $U_{ij}^* = F_{Y_{ij}^* | B_i}(Y_{ij}^* | B_i)$ and $U_{ij} = F_{Y_{ij} | B_i}(Y_{ij} | B_i)$ are the so-called (conditional) probability integral transforms (PITs). The CDF $F_{Y_{ij}^*, Y_{ij} | B_i}$ is thus specified via its margins (given B_i) and the Gaussian copula that glues them together. Note that the margins $F_{Y_{ij}^* | B_i}$ and $F_{Y_{ij} | B_i}$ need not come from the same parametric family, allowing researchers great flexibility in modeling disparate outcomes. The use of Gaussian copula in (4) is also appealing, since it describes dependence in the same way that the multivariate normal distribution does, in addition to its analytical and computational tractability (Song, 2007).

Genest and Nešlehová (2007) demonstrated that copulas directly linking discrete margins are unique only on the Cartesian product of marginal ranges due to non-uniformity of PITs in the discrete case. Although this may seem to be merely a theoretical issue without any practical implications, it has a host of serious consequences that bear directly on dependence modeling of discrete data. For one, common rank-based association measures may now depend on the margins. For another, the range of their possible values may be restricted – severely unnaturally in some cases – rendering interpretations of such measures problematic. Our use of latent variable Y_{ij}^* to describe X_{ij} and our introduction of the copula model at the latent level, manage to sidestep these issues.

2.1. Marginal models

For $j = 1, \dots, n_i$, and $i = 1, \dots, N$, we have

$$\begin{aligned} Y_{ij}^* | B_i &\sim \text{independent } F_{Y_{ij}^* | B_i}, \\ Y_{ij} | B_i &\sim \text{independent } F_{Y_{ij} | B_i}, \end{aligned}$$

where we assume $E(B_i) = 0$ and $\text{var}(B_i) = 1$, with $\text{var}(Y_{ij}^* | B_i) = 1$ (or unit scale), for identifiability reasons. The conditional mean models, given the random effect, are defined by

$$\mu_i^*(B_i) = E(Y_{ij}^* | B_i) = \mu_i^*(\mathbf{z}_{1i}, \boldsymbol{\alpha}) + \lambda_1 B_i, \quad (5)$$

$$\mu_i(B_i) = E(Y_{ij} | B_i) = \mu_i(\mathbf{z}_{2i}, \boldsymbol{\beta}) + \lambda_2 B_i, \quad (6)$$

where \mathbf{z}_{1i} and \mathbf{z}_{2i} are known cluster-level (possibly outcome-specific) covariate vectors, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding unknown regression coefficients, with λ_1 and λ_2 accounting for difference

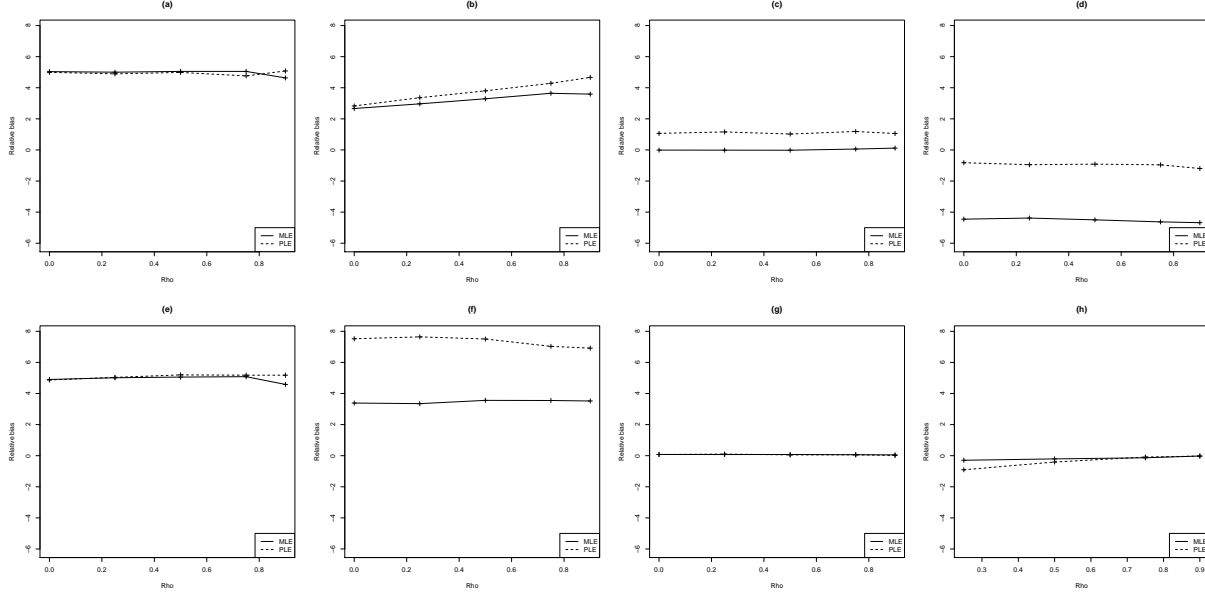


Figure 1: Relative bias of MLEs and PLEs of (a) α_1 , (b) α_2 , (c) λ_1 , (d) β_1 , (e) β_2 , (f) λ_2 , (g) σ , and (h) ρ for robit-normal model.

in scales of Y_{ij}^* and Y_{ij} . We get the marginal means from (5) and (6) as $E(Y_{ij}^*) = \mu_i^*(\mathbf{z}_{1i}, \boldsymbol{\alpha}) = \mathbf{z}_{1i}^\top \boldsymbol{\alpha}$, say, and $E(Y_{ij}) = \mu_i(\mathbf{z}_{2i}, \boldsymbol{\beta}) = \mathbf{z}_{2i}^\top \boldsymbol{\beta}$, say, and the marginal variances as $var(Y_{ij}^*) = 1 + \lambda_1^2$ and $var(Y_{ij}) = \sigma^2 + \lambda_2^2$, where $\sigma^2 = var(Y_{ij}|B_i) > 0$. Note that in the case of binary X_{ij} , the single cutpoint γ can be assumed to be 0, provided an intercept term is included in $\boldsymbol{\alpha}$. In addition, a logistic latent distribution for Y_{ij}^* results in a conditional logistic regression model for X_{ij} , while a normal latent distribution leads to a conditional probit regression model. With normal margins for Y_{ij}^* and Y_{ij} , our model specializes to Najita et al.'s (2010) correlated probit model. A general version of Gueorguieva and Agresti's (2001) model with correlated random effects is discussed in the next section.

2.2. Marginal associations

For the model with shared random effect, the marginal between-subject and within-subject correlations are found to be

$$corr(Y_{ij}^*, Y_{ij'}) = \frac{\lambda_1^2}{1 + \lambda_1^2}, \quad corr(Y_{ij}, Y_{ij'}) = \frac{\lambda_2^2}{\sigma^2 + \lambda_2^2}, \quad (7)$$

$$corr(Y_{ij}^*, Y_{ij}) = \frac{\rho\sigma + \lambda_1\lambda_2}{\sqrt{(1 + \lambda_1^2)(\sigma^2 + \lambda_2^2)}}, \quad (8)$$

$$corr(Y_{ij}^*, Y_{ij'}) = \frac{\lambda_1\lambda_2}{\sqrt{(1 + \lambda_1^2)(\sigma^2 + \lambda_2^2)}}, \quad (9)$$

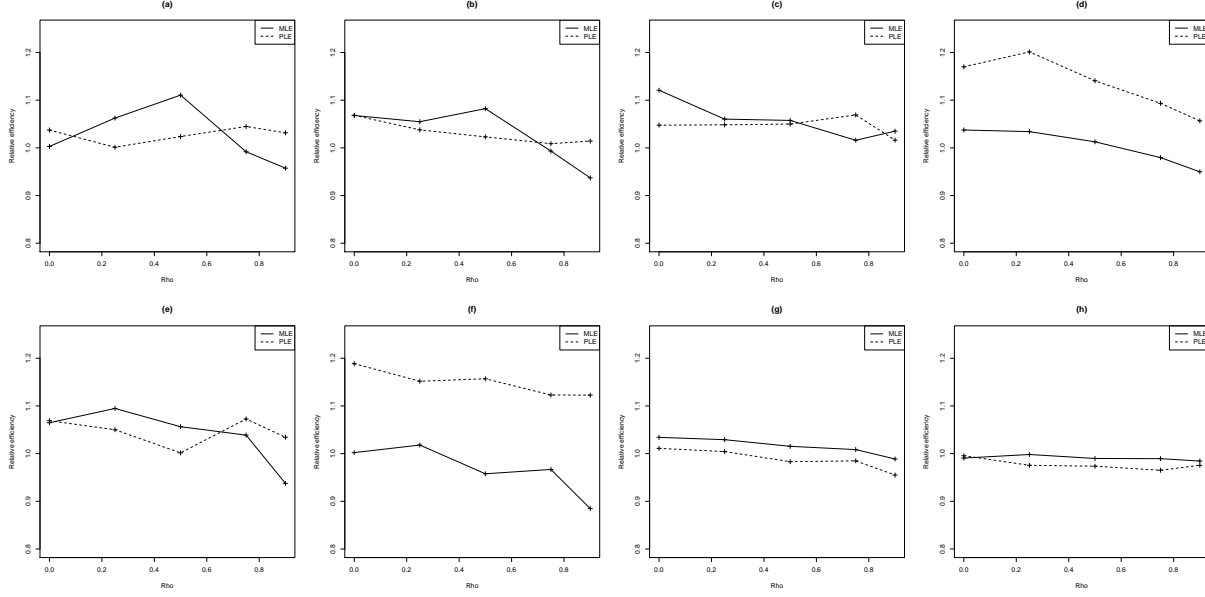


Figure 2: Relative efficiency of MLEs and PLEs of (a) α_1 , (b) α_2 , (c) λ_1 , (d) β_1 , (e) β_2 , (f) λ_2 , (g) σ , and (h) ρ for robit-normal model.

where $\rho = \text{corr}(Y_{ij}^*, Y_{ij}|B_i)$, and where $\text{corr}(Y_{ij}^*, Y_{ij'})$ is the polychoric correlation between X_{ij} and $X_{ij'}$, and $\text{corr}(Y_{ij}^*, Y_{ij})$ is the polyserial correlation between X_{ij} and Y_{ij} . Observe that $\text{corr}(Y_{ij}^*, Y_{ij'}) = \sqrt{\text{corr}(Y_{ij}^*, Y_{ij})\text{corr}(Y_{ij}, Y_{ij'})}$, so that having a shared random effect yields a particularly restrictive association structure for the mixed outcomes from different subjects. This motivates consideration of correlated random effects as follows. For $j = 1, \dots, n_i$, and $i = 1, \dots, N$, we have

$$\mu_i^*(B_{1i}) = \mu_i^*(\mathbf{z}_{1i}, \boldsymbol{\alpha}) + \lambda_1 B_{1i}, \quad (10)$$

$$\mu_i(B_{2i}) = \mu_i(\mathbf{z}_{2i}, \boldsymbol{\beta}) + \lambda_2 B_{2i}, \quad (11)$$

where $E(B_{1i}) = E(B_{2i}) = 0$, $\text{var}(B_{1i}) = \text{var}(B_{2i}) = 1$, and $\text{cov}(B_{1i}, B_{2i}) = \text{corr}(B_{1i}, B_{2i}) = \rho_B$. Expressions (7) remain unchanged; correlations (8)–(9) are modified accordingly as

$$\text{corr}(Y_{ij}^*, Y_{ij}) = \frac{\rho\sigma + \rho_B\lambda_1\lambda_2}{\sqrt{(1 + \lambda_1^2)(\sigma^2 + \lambda_2^2)}}, \quad (12)$$

$$\text{corr}(Y_{ij}^*, Y_{ij'}) = \frac{\rho_B\lambda_1\lambda_2}{\sqrt{(1 + \lambda_1^2)(\sigma^2 + \lambda_2^2)}}, \quad (13)$$

where we again assumed $\text{var}(Y_{ij}^*|B_{1i}, B_{2i}) = 1$ and $\text{var}(Y_{ij}|B_{1i}, B_{2i}) = \sigma^2$. Noting that $\text{corr}(Y_{ij}^*, Y_{ij'}) = \rho_B\sqrt{\text{corr}(Y_{ij}^*, Y_{ij})\text{corr}(Y_{ij}, Y_{ij'})}$, we see that a more flexible association structure is possible

with a model with correlated random effects. Instead of the usual joint normality assumption for B_{1i} and B_{2i} as in Gueorguieva and Agresti (2001), it is possible to similarly use the Gaussian copula to build their joint distribution as in Lin et al. (2010). This affords flexibility in specifying the marginal distributions of B_{1i} and B_{2i} ; for example, a bridge margin may be assumed for B_{1i} to facilitate interpretability of marginal effects with a logistic regression model for binary outcome X_{ij} (i.e., a logistic latent distribution for Y_{ij}^*).

Note that correlations between outcomes are not modeled directly in our approach; instead, correlations are incorporated for the normal scores, which are transformations of the original variables (or latent variables in the case of discrete data assumed to have latent structure). For example, ρ_Z in (4) is used as proxy for ρ ; it can be shown that $\rho \leq |\rho_Z|$ (Klaassen and Wellner, 1997), which can then be used to bound marginal correlations (8) and (13). Alternatively, a piecewise-linear approximation may be used to recover ρ from ρ_Z (Kugiumtzis and Borasenta, 2010). We can also use nonparametric rank-based measures like Kendall's tau to gauge associations, since they are invariant to monotonic transformations. For the Gaussian copula model above with shared random effect, we get the conditional Kendall's tau as $\tau(Y_{ij}^*, Y_{ij}|B_i) = \tau(Q_{ij}^*, Q_{ij}|B_i) = 2 \arcsin(\rho_Z)/\pi$.

3. Likelihood estimation

Let $\{x_{ij}, y_{ij}, \mathbf{z}_{1i}, \mathbf{z}_{2i}\}$ denote the observed data, for $j = 1, \dots, n_i$, and $i = 1, \dots, N$. Assuming a shared random effect model, with Θ as the parameter vector containing the regression coefficients α and β , the correlation ρ_Z , and respective parameters θ_1 and θ_2 of (conditional) margins $F_{Y_{ij}^*|B_i}$ and $F_{Y_{ij}|B_i}$, the likelihood contribution of cluster i is given by

$$L_i(\Theta) = \int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} f_{X_{ij}, Y_{ij}|B_i}(x_{ij}, y_{ij}|b_i) f_{B_i}(b_i) db_i, \quad (14)$$

so that the likelihood function becomes $L(\Theta) = \prod_{i=1}^N L_i(\Theta)$. The log-likelihood function is then $\ell(\Theta) = \log L(\Theta) = \sum_{i=1}^N \log L_i(\Theta)$. In the mixed binary-continuous case, we get a simple expression for $f_{X_{ij}, Y_{ij}|B_i}$ as follows:

$$f_{X_{ij}, Y_{ij}|B_i}(x_{ij}, y_{ij}|b_i) = \begin{cases} \frac{1}{\sigma} \phi\left(\frac{y_{ij} - \mathbf{z}_{2i}^\top \beta}{\sigma}\right) \Phi\left(\frac{\frac{\rho_Z}{\sigma}(y_{ij} - \mathbf{z}_{2i}^\top \beta) - \Phi^{-1}(u_{ij}^*)}{\sqrt{1 - \rho_Z^2}}\right) & , \text{ if } x_{ij} = 1 \\ \frac{1}{\sigma} \phi\left(\frac{y_{ij} - \mathbf{z}_{2i}^\top \beta}{\sigma}\right) \Phi\left(\frac{\Phi^{-1}(u_{ij}^*) - \frac{\rho_Z}{\sigma}(y_{ij} - \mathbf{z}_{2i}^\top \beta)}{\sqrt{1 - \rho_Z^2}}\right) & , \text{ if } x_{ij} = 0 \end{cases},$$

where ϕ is the standard normal PDF, and $u_{ij}^* = P(Y_{ij}^* < 0 | B_i = b_i) = F_{Y_{ij}^*|B_i}(0 | b_i; \mathbf{z}_{1i}, \alpha, \theta_1) = \int_{-\infty}^0 f_{Y_{ij}^*|B_i}(y^* | b_i) dy^*$.

Putting $s(\Theta) = \partial \ell(\Theta) / \partial \Theta$ as the score function and $h(\Theta) = \partial^2 \ell(\Theta) / \partial \Theta \partial \Theta^\top$ as the Hessian matrix, the MLE $\hat{\Theta}$ is obtained by solving the likelihood equations $s(\Theta) = \mathbf{0}$ iteratively via a Newton-type algorithm. It can be easily verified that $\hat{\Theta}$ is consistent and asymptotically multivariate normal with mean Θ and covariance matrix given by the inverse of the Fisher information matrix $E\{-h(\Theta)\} = E\{s(\Theta)s^\top(\Theta)\}$. Standard errors (SEs) for $\hat{\Theta}$ are calculated from diagonals of $\{s(\hat{\Theta})s^\top(\hat{\Theta})\}^{-1}$ or $-h^{-1}(\hat{\Theta})$, provided either matrix is invertible.

Table 1: Summary of EG mice data.

Dose (g/kg)	Number of dams	Number of live fetuses	Malformations		Weight (g)	
			Number	Percent	Mean	SD
0	25	297	1	0.3	0.972	0.098
0.75	24	276	26	9.42	0.877	0.104
1.5	22	229	89	38.86	0.764	0.107
3	23	226	129	57.08	0.704	0.124

4. Simulation study

In this section, we consider a conditional joint model assembled from a normal margin for continuous outcome Y_{ij} and a t -margin for latent variable Y_{ij}^* underlying binary outcome $X_{ij} = I(Y_{ij}^* \geq 0)$, $j = 1, \dots, n_i$, $i = 1, \dots, N$. Specifically, given random effect $B_i \sim n(0, 1)$ shared by measurements in cluster i , we have

$$Y_{ij}^*|B_i \sim \text{independent } t_\nu(\mu_i^*(B_i), 1), \quad (15)$$

$$Y_{ij}|B_i \sim \text{independent } n(\mu_i(B_i), \sigma^2), \quad (16)$$

where $\mu_i^*(B_i) = \alpha_1 + \alpha_2 z_{1i} + \lambda_1 B_i$ and $\mu_i(B_i) = \beta_1 + \beta_2 z_{2i} + \lambda_2 B_i$, with outcome-specific cluster-level covariates z_{1i} and z_{2i} . The model, called robit-normal model by de Leon and Wu (2011), yields a robit regression model (Liu, 2004) for X_{ij} , a robust alternative to the logistic and probit models. With $\nu \approx 7$, the latent mixed model is equivalent to a mixed effects logistic regression model for X_{ij} ; with large ν , it yields a mixed effects probit regression model.

For the simulations, a total $R = 1000$ repeated samples for full maximum likelihood estimation and $R = 500$ repeated samples for profile likelihood estimation are generated with $N = 200$ clusters, each with varying sizes from $n_i = 1, \dots, 10$, generated randomly via a binomial distribution. Covariates z_{1i} and z_{2i} are also generated using a uniform distribution over $(-1, 1)$. The following parameter configuration was considered: $\alpha_1 = \beta_1 = \lambda_1 = \lambda_2 = \sigma = 1$, and $\alpha_2 = \beta_2 = 2$, with varying correlation $\rho_Z \in \{0, 0.25, 0.5, 0.75, 0.9\}$. The relative bias RB = (mean of estimates – true value)/true value $\times 100$, and relative efficiency RE = mean of SEs/empirical SD of estimates are considered. To avoid problems associated with estimating the degrees of freedom ν , we employ the method of profile likelihood (Song et al., 2007). The method entails maximizing the profile log-likelihood $\ell_\nu(\Theta_{(-\nu)}) = \ell(\Theta_{(-\nu)}; \nu)$ at fixed grid points $\nu \in (2, B]$, for some suitably large constant B , where $\Theta_{(-\nu)}$ is Θ with ν removed; in the simulations, we fixed $B = 8$. Our estimates $\hat{\Theta}_{(-\nu)} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2, \hat{\rho}_Z)^\top$ correspond to $\hat{\nu}$ at which $\ell_\nu(\Theta_{(-\nu)})$ is maximum on $(2, B]$. This method is easy to implement using dense grid points on $(2, B]$, and is more computationally efficient than full maximum likelihood estimation. Asymptotic properties analogous to those for MLEs can be similarly established. We used PROC NL MIXED in SAS to implement the estimation.

Figure 1 plots relative biases of MLEs and PLEs as functions of $\rho_Z = 0.25, 0.5, 0.75, 0.9$; note that the relative bias of the MLE and PLE of ρ_Z is undefined for $\rho_Z = 0$. Figure 2 plots the corresponding relative efficiencies of the MLEs and PLEs. From Figure 1, we can see that both MLEs and PLEs are relatively unbiased, with all plots close to 0. Figure 2 displays relative efficiencies that are generally close to 1, indicating that both MLEs and PLEs have SEs that reflect the estimates' true sampling variability.

Table 2: PLEs and their SEs and z-values for EG mice data.

Parameter	Est	SE	z
<i>Fetal malformation</i>			
α_1	-1.2522	0.1847	-6.78
α_2	2.9594	0.0893	33.15
λ_1	0.3639	0.0882	4.13
ν	6.5	-	-
<i>Fetal weight</i>			
β_1	0.9377	0.0141	66.73
β_2	-0.0828	0.0077	-10.7
λ_2	0.0907	0.0059	15.17
σ	0.0749	0.0018	42.61
<i>Association</i>			
ρ_Z	-0.1326	0.0552	-2.4
τ	-0.0847	0.033	-2.57

5. Example: EG mice data

In this section, we adopt the robit-normal mixed model in Section 4 to analyze the EG mice data described in Section 1. A summary of the data is displayed in Table 1. It is apparent that fetal weight decreases with increasing dose on average, with the mean weight ranging from 0.972g at dose 0 to 0.704g at the highest dose. Similarly, the proportion of malformed fetuses monotonically increases with dose, from about 0.3% at dose 0 up to about 58% at dose 3g/kg/day.

In our analysis, we assume litters (i.e., dams) are independent. Two outcomes which are believed to be indicative of toxicity are considered: X_{ij} , a binary outcome indicating presence ($X_{ij} = 1$) or absence ($X_{ij} = 0$) of fetal malformations; and Y_{ij} , a continuous outcome representing fetal weight, for fetus $j = 1, \dots, n_i$ in dam $i = 1, \dots, 94$. The primary interest is the simultaneous effects of the common covariate dose $z_i = 0, 0.75, 1.5, 3$ g/kg, on both outcomes. We assume a latent variable $Y_{ij}^*|B_i \sim t_\nu(\mu_i^*(B_i), 1)$, such that $X_{ij} = I(Y_{ij}^* \geq 0)$, and consider (conditional) linear mixed models relating conditional means $\mu_i^*(B_i) = \alpha_1 + \alpha_2 z_i + \lambda_1 B_i$ and $\mu_i(B_i) = \beta_1 + \beta_2 z_i + \lambda_2 B_i$, given $B_i \sim n(0, 1)$. The choice of a robit mixed model for X_{ij} results in analysis that is robust to presence of outliers (Liu, 2004). It also provides a general approach to binary regression modelling since a robit model approximates both logit and probit models.

The resulting PLEs for the robit-normal mixed model for the EG mice data are reported in Table 2. We can see that $\hat{\alpha}_2 = 2.9594$, in the robit regression of fetal malformation on dose, indicates that the conditional malformation probability increases with increasing dose. Similarly, $\hat{\beta}_2 = -0.0828$, in the linear mixed model of fetal weight on dose, suggests that the mean fetal weight decreases with increasing dose. The dose coefficient for fetal malformation and weight are both statistically significant.

The estimated correlation $\hat{\rho}_Z = -0.1326$, suggests that a slightly negative association exists between fetal weight and the latent variable underlying fetal malformation. Kugiumtzis and Bora-Senta's (2010) piecewise linear approximation method yields an estimated conditional polyserial

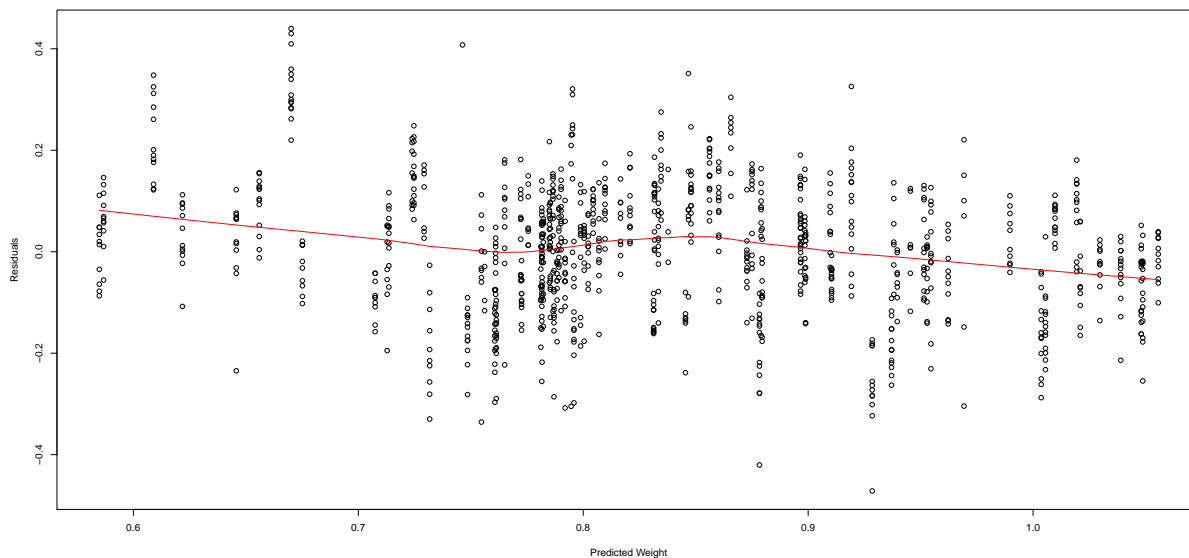


Figure 3: Plot of weight residual against predicted fetal weight for robit-normal mixed model.

correlation between fetal malformation and fetal weight of $\hat{\rho} = -0.1282$. The estimated Kendall's tau for conditional association $\hat{\tau}(Y_{ij}^*, Y_{ij}|B_i)$ is also displayed along with its SE obtained by the delta method.

A loess-smoothed plot of fetal weight residuals against predicted fetal weight is shown in Figure 3. The plot displays no apparent trend, which suggests that the model provides an adequate fit for the data. Catalano and Ryan (1992) and Fitzmaurice and Laird (1995), among others, have both previously analyzed the data using various mixed-outcome regression models based on factorization approaches which accounted for clustering. While not all parameters are comparable because of the different directions of conditioning and link functions used, the dose coefficients as well as the associations may be contrasted. For example, Catalano and Ryan (1992) used probit models for the binary malformation outcome while Fitzmaurice and Laird (1995) adopted a logistic regression model. The former, on the one hand, gave respective estimates of the intercept and dose coefficient as 0.963 and -0.0952 , with respective SEs 0.012 and 0.00832 for their fetal weight model, and respective estimates of the intercept and dose coefficient for their malformation model as -2.07 and 0.831, with respective SEs 0.145 and 0.109; the inter-fetus correlation is 0.48. The latter, on the other hand, reported respective estimates of the intercept and dose coefficient for their fetal weight model as 0.9537 and -0.089 , with respective SEs 0.0123 and 0.0079; for their fetal malformation model, their estimates of the intercept and dose coefficient are, respectively, -2.993 and 1.1984, with respective SEs 0.3102 and 0.1461. Their estimates of the intra-litter correlations for fetal weight and malformation are 0.4521 and 0.2485, respectively. These are generally in agreement with our estimates in Table 2. Comparing our results against those reported by Lin et al. (2010) and Gueorguieva and Agresti (2001), it appears that while the estimates are generally quite close, the SEs of our estimates are slightly smaller than theirs. The closeness of estimates is not surprising since $\hat{\nu} = 6.5 \approx 7$, suggesting that the t -latent margin approximates a logit model.

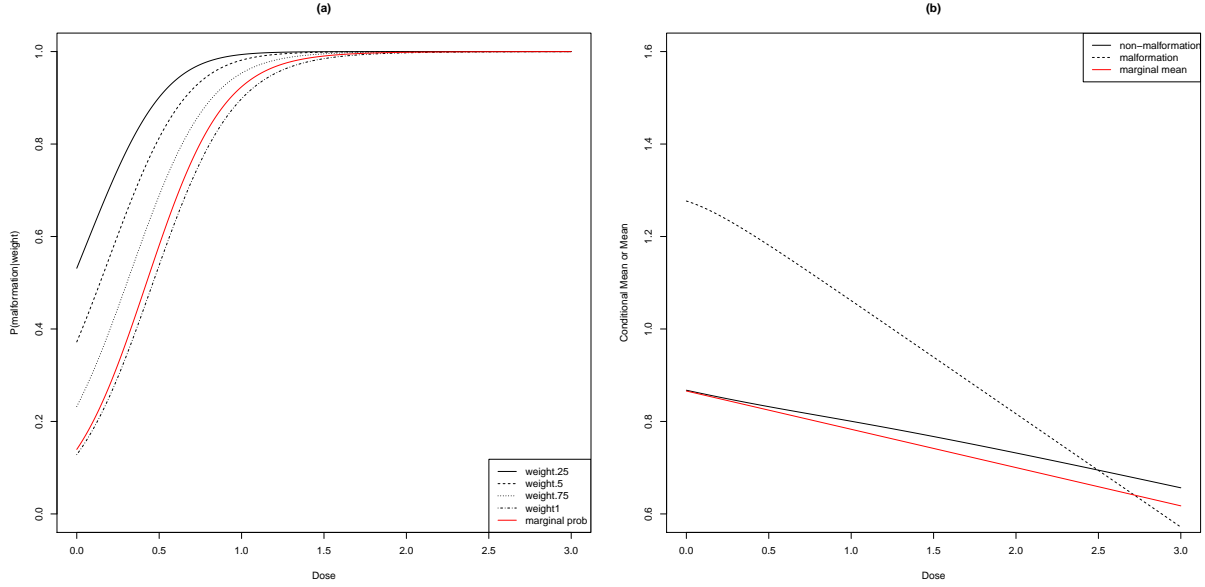


Figure 4: Plots of estimated conditional (a) fetal malformation probability (given fetal weight) and (b) mean fetal weight (given malformation indicator), as functions of dose. The estimated marginal fetal malformation probability and marginal mean fetal weight, both as functions of dose, are also shown for comparison.

Because we have a joint model, it is possible to obtain conditional interpretations for one response variable given the other. For the EG mice data, we can obtain the conditional distribution of fetal malformation as a function of age, given fetal weight, or vice versa. To see this, we obtained the conditional probability of malformation at fixed doses, given fetal weight. From Section 3, we have

$$P(X_{ij} = x | Y_{ij} = y; z_i) = \begin{cases} \int_{-\infty}^{+\infty} \Phi \left(\frac{\frac{\rho_Z}{\sigma}(y - \mathbf{z}_i^\top \boldsymbol{\beta}) - \Phi^{-1}(u_{ij}^*)}{\sqrt{1 - \rho_Z^2}} \right) \phi(b) db & , \text{ if } x = 1 \\ \int_{-\infty}^{+\infty} \Phi \left(\frac{\Phi^{-1}(u_{ij}^*) - \frac{\rho_Z}{\sigma}(y - \mathbf{z}_i^\top \boldsymbol{\beta})}{\sqrt{1 - \rho_Z^2}} \right) \phi(b) db & , \text{ if } x = 0 \end{cases} ,$$

where $u_{ij}^* = F_{Y_{ij}^* | B_i}(0 | b_i; \mathbf{z}_i, \boldsymbol{\alpha}, \nu)$, $\mathbf{z}_i = (1, z_i)^\top$, and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$; similarly, we get conditional mean fetal weight, as a function of dose, as

$$E(Y_{ij} | X_{ij} = x; z_i) = \begin{cases} \int_{-\infty}^{+\infty} \left(E(Y_{ij} | b) + \frac{\rho_Z \sigma}{1 - u_{ij}^*} \phi\{\Phi^{-1}(u_{ij}^*)\} \right) \phi(b) db & , \text{ if } x = 1 \\ \int_{-\infty}^{+\infty} \left(E(Y_{ij} | b) - \frac{\rho_Z \sigma}{u_{ij}^*} \phi\{\Phi^{-1}(u_{ij}^*)\} \right) \phi(b) db & , \text{ if } x = 1 \end{cases} .$$

We plotted the estimated conditional malformation probability as a function of fetal weight $y = 0.25, 0.5, 0.75, 1\text{g}$ at each dose level $z = 0, 0.75, 1.5, 3\text{g/kg}$, in Figure 5(a), and the conditional

mean fetal weight, as a function of dose, given malformation indicator $x = 0$ or $x = 1$, in Figure 5(b). Observe that for given fetal weight, the estimated conditional probability of malformation increases with the dose; we also find that larger fetuses are less likely to have malformations. Given the malformation indicator, the conditional mean fetal weight decreases with dose. For comparisons, we include in Figure 5 the marginal probability of malformation and marginal mean fetal weight (shown as solid red plots), as functions of dose.

6. Discussion

In this article, we developed a Gaussian copula-based regression model for mixed bivariate discrete and continuous outcomes by introducing random effects to incorporate within-cluster associations. The marginal regression models are specified using generalized linear models linking the outcomes' marginal means to covariates.

We paid particular attention to the robit-normal mixed model for binary and normal endpoints, where a t -latent distribution is used to formulate the marginal distribution of the binary endpoint. The model for the binary outcome corresponds to so-called robit regression, a robust alternative to and extension of logit and probit models. The resulting regression parameters α and β have marginal interpretations, which are often the main interest in applications. By introducing cluster-specific random effects, our model can efficiently handle within-cluster associations between mixed/same-type outcomes on the same/different fetuses. Adopting a latent variable formulation of the discrete outcome and using Gaussian copula to indirectly specify the joint model at the latent level, we are able to sidestep complications arising from direct application of copula to discrete data. The correlation can be interpreted as proxy for the polyserial correlation between continuous and discrete endpoints.

Simulation results suggest that the (both full and profile) likelihood estimation is able to estimate parameters well and their SEs reflect true sampling variability. As an illustration of our approach, we applied the robit-normal mixed model to the EG mice data. Our results were similar to those in Lin et al. (2010) and Gueorguieva and Agresti (2001), albeit ours yielded slightly smaller SEs. As in Lin et al. (2010), our methodology can be easily implemented using standard statistical software/package; in particular, we used PROC NLMIXED in SAS in our simulations and analysis.

Acknowledgment

This work was partially supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- [1] Catalano, P. J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, **16**, 883–900.
- [2] Catalano, P. J., and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651–658.
- [3] de Leon, A. R., and Carrière, K. C. (2010). Mixed-outcome data. In *Encyclopedia of Biopharmaceutical Statistics*, vol. 2, Chow, S-C. (ed.), pp. 817–822, Chapman & Hall.

- [4] de Leon, A. R., and Carrière, K. C. (2007). General mixed-data model: extension of general location and grouped continuous models. *Canadian Journal of Statistics*, **35**, 533–548.
- [5] de Leon, A. R., and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, **30**, 175–185.
- [6] de Leon, A. R., and Zhu, Y. (2008). ANOVA extensions for mixed discrete and continuous data. *Computational Statistics & Data Analysis*, **52**, 2218–2227.
- [7] Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., and Bijmens, L. (2009). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, **27**, 4408–4427.
- [8] Faes, C., Geys, H., and Catalano, P. (2009). Joint models for continuous and discrete longitudinal data. In *Longitudinal Data Analysis*, Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (eds.), pp. 327–348, Chapman & Hall.
- [9] Faes, C., Geys, H., Aerts, M., Molenberghs, G., and Catalano, P. (2004). Modelling combined continuous and ordinal outcomes in a clustered setting. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 515–530.
- [10] Fitzmaurice, G. M., and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, **53**, 110–122.
- [11] Fitzmaurice, G. M., and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845–852.
- [12] Genest, C., and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin*, **37**, 475–515.
- [13] George, E. O., Armstrong, D., Catalano, P. J., and Srivastava, D. K. (2007). Regression models for analyzing clustered binary and continuous outcomes under an assumption of exchangeability. *Journal of Statistical Planning and Inference*, **137**, 3462–3474.
- [14] Gueorguieva, R. V., and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, **25**, 1307–1322.
- [15] Gueorguieva, R. V., and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, **96**, 1102–1112.
- [16] Hirakawa, A. (2011). An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. To appear in *Statistics in Medicine*.
- [17] Klaassen, C. A. J., and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, **3(1)** 55–77.
- [18] Kugiumtzis, D., and Bora-Senta, E. (2010). Normal correlation coefficient of non-normal variables using piece-wise linear approximation. *Computational Statistics*, **25**, 645–662.
- [19] Lin, L., Bandyopadhyay, D., Lipsitz, S. R., and Sinha, D. (2010). Association models for clustered data with binary and continuous responses. *Biometrics*, **66** 287–293.

- [20] Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*, Gelman, A., and Meng, X-L. (eds.), Wiley, 227–238.
- [21] McCulloch, C. (2007). Joint modeling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, **17**, 1–21.
- [22] Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.
- [23] Najita, J. S., Li, Y., and Catalano, P. J. (2009). A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society-C*, **58**, 555–573.
- [24] Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicological Applications in Pharmacology*, **81**, 113–127.
- [25] Regan, M. M., and Catalano, P. J. (2002). Combined continuous and discrete outcomes. In *Topics in Modelling of Clustered Data*, Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (eds.), pp. 233–261, Chapman & Hall.
- [26] Regan, M. M., and Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, **55**, 760–768.
- [27] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- [28] Song, P. X-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.
- [29] Song, P. X-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copula. *Biometrics*, **65**, 60–68.
- [30] Song, P. X.-K., Zhang, P., and Qu, A. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica*, **17**, 929–943.
- [31] Teixeira-Pinto, A., and Normand, S.-L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in Medicine*, **28**, 1753–1773.
- [32] Zimmer, D. M., and Trivedi, P. K. (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business and Economic Statistics*, **24**, 63–76.