JIRSS (Year)
Vol. xx, No. x, pp xx-xx
DOI:00000000000000000000000000000000

Conditional Dependence in Longitudinal Data Analysis

Mahmoud Torabi [1] ✉, **Alexander R. de Leon** [2]

[1]Departments of Community Health Sciences & Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3E 0W3.
[2]Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4.

Abstract. Mixed models are widely used to analyze longitudinal data. In their conventional formulation as linear mixed models (LMMs) and generalized LMMs (GLMMs), a commonly indispensable assumption in settings involving longitudinal non-Gaussian data is that the longitudinal observations from subjects are conditionally independent, given subject-specific random effects. Although conventional Gaussian LMMs are able to incorporate conditional dependence of longitudinal observations, they require that the data are, or some transformation of them is, Gaussian, a serious limitation in a wide variety of practical applications. Here, we introduce the class of Gaussian copula conditional regression models (GCCRMs) as flexible alternatives to conventional LMMs and GLMMs. One advantage of GCCRMs is that they extend conventional LMMs and GLMMs in a way that reduces to conventional LMMs, when the data are Gaussian, and to conventional GLMMs, when conditional independence is assumed. We implement likelihood analysis of GCCRMs using existing software and statistical packages and evaluate the finite-sample performance of maximum likelihood estimates for GCCRM empirically via simulations vis-à-vis the 'naive' likelihood analysis that incorrectly assumes conditionally independent longitudinal data. Our results show that the 'naive' analysis yields estimates with possibly severe bias and incorrect standard errors, leading to misleading inferences. We use bolus count data on patients' controlled analgesia comparing dosing regimes and data on serum creatinine from a renal graft study to illustrate the applications of GCCRMs.

Keywords. Exponential family, Gaussian copula, Marginal distribution, Maximum likelihood estimation, Random effects

Corresponding Author(✉): Mahmoud Torabi (Mahmoud.Torabi@umanitoba.ca), Alexander R. de Leon (adeleon@ucalgary.ca)

MSC: 62H12, 62J12.

1 Introduction

A common example of correlated data arise frequently in longitudinal studies, where an outcome or outcomes are measured repeatedly over a period of time (i.e., longitudinally) from the same subjects. As such, the resulting observations are correlated, as when weekly blood pressure readings are obtained from cohorts of treated and untreated patients. Such longitudinal data, especially in health and medical research (Brown and Prescott, 2015; Magezi, 2015), are often analyzed using conventional (i.e., Gaussian) linear mixed models (LMMs) (Searle, Casella, and McCulloch, 2006) and generalized LMMs (GLMMs) (McCulloch, Searle, and Neuhaus, 2008) to account for population heterogeneity, over- or under-dispersion, and within-subject correlations in the data via the inclusion of subject-specific random effects. They also enable the borrowing of information across different subjects for subject-specific predictions, as opposed to population-averaged predictions based on marginal models.

Let $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iT_i})^\top$ be the vector of longitudinal observations on a single outcome for subject $i(= 1, \cdots, n)$ up to and including time $T_i(\geq 1)$. Suppose that Y_{it} is amenable to be modelled by a LMM (e.g., it is not strictly positive, such as a survival endpoint) but it may not follow the Gaussian distribution. Suppose further that normalizing Y_{it} via transformations is not attractive for reasons of interpretability. Any longitudinal analysis of the data needs to account for the marginal longitudinal dependence in $\boldsymbol{Y}_i$ as captured by the marginal longitudinal correlations $corr(Y_{it}, Y_{it'})$ between any pair of longitudinal observations Y_{it} and $Y_{it'}$, with $t \neq t'$.

The default approach in practice relies on a conventional LMM for Y_{it} defined as

$$Y_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{B}_i + \epsilon_{it}, \tag{1.1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients for the vector $\boldsymbol{x}_{it}$ of covariates (possibly including time t), $\boldsymbol{B}_i$ is the vector of subject-specific Gaussian random effects (possibly containing random slopes of time) with the design vector of $\boldsymbol{z}_i$, and $\epsilon_{i1}, \cdots, \epsilon_{iT_i}$ are either independent or correlated Gaussian residual errors. The Gaussian random effects and Gaussian residual errors together imply that $\boldsymbol{Y}_i$ is, conditionally and marginally, Gaussian as well.

Note that conventional LMMs rely on the assumption that the outcome, or some transformation of it, follows a Gaussian distribution. This has been shown to be very restrictive in many applications and working with a transformed outcome often engenders issues concerning interpretability. While conventional LMMs can incorporate conditional dependence in the model by allowing for correlated Gaussian residual errors, the absence of a general multivariate non-Gaussian model similar to the multivariate Gaussian distribution for non-Gaussian $\boldsymbol{Y}_i$ often necessitates assuming conditional independence of the non-Gaussian longitudinal observations $Y_{i1}, \cdots, Y_{iT_i}$, given the random effects $\boldsymbol{B}_i$. This is tantamount to assuming that the marginal longitudinal

dependence between successive observations is completely explained by $\boldsymbol{B}_i$, a very strong and oftentimes invalid assumption in practice, since the observations may be intrinsically correlated over time, as when $Y_{i1}, \cdots, Y_{iT_i}$ are biologically linked (Das et al., 2013).

In non-Gaussian settings involving longitudinal observations $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iT_i})^\top$ that do not fit the LMM (e.g., a survival endpoint or a binary/count outcome), GLMMs are widely and typically adopted. Given a non-identity link $g(\cdot)$, a conventional GLMM for Y_{it} is given by

$$g[E(Y_{it}|\boldsymbol{B}_i)] \quad = \quad \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{B}_i, \tag{1.2}$$

where Y_{it}, given $\boldsymbol{B}_i$, has some parametric non-Gaussian distribution (e.g., gamma, binomial, Poisson). Conditional independence of $Y_{i1}, \cdots, Y_{iT_i}$ conveniently allows for the construction of a conditional joint model (i.e., conditional distribution of $\boldsymbol{Y}_i$) as simply the product of the conditional marginal models (i.e., conditional marginal distributions of $Y_{i1}, \cdots, Y_{iT_i}$). Such an assumption, which may not hold in practice, becomes indispensable because convenient multivariate generalizations of the gamma, Bernoulli and Poisson distributions, among many other non-Gaussian models, are likewise generally unavailable. Conditional dependence in this situation may be accommodated by adopting a factorization or conditioning approach, as in so-called transition models; however, except in the case of conventional LMMs, the use of transition models for non-Gaussian data are not without its own complications.

Our objective in this paper is to explore the impact on estimation in LMMs and GLMMs when conditional independence of longitudinal non-Gaussian observations is incorrectly assumed in the analysis. Wu, de Leon and Withanage (2013) and Wu and de Leon (2014) addressed the related issue of conditional dependence among different outcomes in the context of joint modelling of correlated data on multiple disparate non-Gaussian outcomes (e.g., mixed binary and continuous outcomes); to our knowledge, a similar investigation has yet to be carried out in relation to conditional dependence among longitudinal observations on a single non-Gaussian outcome. Work by Masarotto and Varin (2012) on Gaussian copula marginal models and by Wu and de Leon (2014) (see also de Leon and Wu, 2011) on Gaussian copula mixed models together provide an especially appropriately convenient framework for directly incorporating conditional dependence in the analysis, thus allowing us to isolate the effect of this particular model mis-specification on the analysis while still assuming that the outcome's mean model is correctly specified.

The paper is organized as follows. We first review conventional LMMs and GLMMs in Section 2, highlighting their inadequacies as far as accounting for conditional dependence in the analysis is concerned. Section 3 briefly reviews copulas and the way they are employed in the construction of (conditional) joint models based on the (conditional) marginal models, with particular attention given to the Gaussian copula on which Masarotto and Varin's (2012) marginal models and Wu and de Leon's (2014) mixed models are based. A brief discussion of marginal longitudinal dependence in terms of correlations among longitudinal observations is included as well; likelihood

analysis for the resulting model is likewise discussed. In Section 4, empirical results of simulation studies on the effect on the bias and efficiency of likelihood-based estimates of mis-specifying the joint model by ignoring conditional dependence are reported. Section 5 illustrates the methodology on longitudinal data from two studies (a study on bolus count data on patient controlled analgesia comparing two different dosing regimes (Weiss, 2005) and another study on renal graft failure (Fieuws and Verbeke, 2008). Section 6 concludes the paper.

2 Conditional Dependence in Longitudinal Data

Marginal dependence is often modeled in conventional LMMs and GLMMs by assuming conditional independence of the correlated observations, given subject-specific random effects. We discuss these models' flexibility in accounting for conditional dependence of Gaussian and non-Gaussian longitudinal data in what follows.

2.1 LMMs for longitudinal Gaussian data

For longitudinal observation Y_{it} from subject $i(= 1, \cdots, n)$ at time $t(= 1, \cdots, T_i)$ described by LMM (1.1) with Gaussian residual errors and Gaussian random effects, the vector $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iT_i})^\top$ of longitudinal observations from subject i has a T_i-dimensional Gaussian marginal distribution with marginal covariance matrix

$$\boldsymbol{V}_i \;=\; \boldsymbol{Z}_i cov(\boldsymbol{B}_i)\boldsymbol{Z}_i^\top + cov(\boldsymbol{\epsilon}_i) \;=\; \boldsymbol{Z}_i \boldsymbol{\Sigma}_B \boldsymbol{Z}_i^\top + \boldsymbol{\Sigma}_i, \tag{2.1}$$

with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{iT_i})^\top$ and $\boldsymbol{Z}_i = \boldsymbol{I}_{T_i} \otimes z_i^\top$ the Kronecker product of the T_i-dimensional identity matrix $\boldsymbol{I}_{T_i}$ and $z_i^\top$, and where $var(\epsilon_{it}) = var(\epsilon_{i't})$, for all $i \neq i'$ such that $t \leq \min(T_i, T_{i'})$, so that we indexed $\boldsymbol{\Sigma}_i$ by i only because its dimension depends on T_i. The decomposition of $\boldsymbol{V}_i$ in (2.1) implies that the marginal longitudinal dependence of $Y_{i1}, \cdots, Y_{iT_i}$ can arise from either $\boldsymbol{\epsilon}_i$ alone (i.e., conditionally, given $\boldsymbol{B}_i$) or $\boldsymbol{B}_i$ alone (with inclusion of random slopes of time), or induced by both.

Assuming $\boldsymbol{\Sigma}_i$ is diagonal is equivalent to assuming that the Gaussian observations $Y_{i1}, \cdots, Y_{iT_i}$ are conditionally independent, given $\boldsymbol{B}_i$. That is, $cov(Y_{it}, Y_{it'}) = z_i^\top \boldsymbol{\Sigma}_B z_i$, so that the marginal longitudinal dependence between Y_{it} and $Y_{it'}$, for $t < t'$, is completely characterized by $\boldsymbol{B}_i$. Inclusion of random slopes of time implies $z_i \equiv z_{it}$, so that $\boldsymbol{Z}_i = diag(z_{i1}^\top, \cdots, z_{iT_i}^\top)$; it follows that $cov(Y_{it}, Y_{it'}) = z_{it}^\top \boldsymbol{\Sigma}_B z_{it'}$, and the marginal longitudinal correlation varies with time, a desirable dependence structure for longitudinal (i.e., temporally ordered) data. Note, however, that such a marginal longitudinal dependence structure does not account for conditional dependence, since the latter may not even be longitudinal in nature.

Conditional dependence of $Y_{i1}, \cdots, Y_{iT_i}$ is conveniently accommodated by a non-diagonal specification for $\boldsymbol{\Sigma}_i$. For example, an AR(1) specification for $\boldsymbol{\Sigma}_i$ corresponds to a stationary first-order autoregressive model for $\boldsymbol{Y}_i$ (Molenberghs and Verbeke, 2005);

this follows easily from the Gaussian distribution's closure properties under conditionalization and marginalization. Such a model assumes that the conditional dependence of $Y_{i1}, \cdots, Y_{iT_i}$ is longitudinal as well, in which case, inclusion of random slopes of time in $\boldsymbol{B}_i$ may not be necessary. It is likewise possible to have, say, a compound symmetric (i.e., exchangeable) specification for $\boldsymbol{\Sigma}_i$ — a non-longitudinal conditional dependence structure for $Y_{i1}, \cdots, Y_{iT_i}$ — thus implying that the marginal longitudinal dependence of $Y_{i1}, \cdots, Y_{iT_i}$ is completely characterized by $\boldsymbol{B}_i$, which necessitates inclusion of random slopes of time in order for the resulting marginal longitudinal correlations to be time-varying.

Conventional LMMs provide a straightforward way of accommodating conditional dependence in the longitudinal data, provided the longitudinal observations are, or some transformation of them is, Gaussian. The lack of similar multivariate analogues of common non-Gaussian distributions presents a difficulty for non-Gaussian longitudinal data.

2.2 GLMMs for longitudinal non-Gaussian data

If the longitudinal observations $Y_{i1}, \cdots, Y_{iT_i}$ are non-Gaussian (e.g., binary, count or survival outcome), then no standard multivariate model exists that mimics the closure properties of the multivariate Gaussian distribution. As such, given the respective conditional marginal probability density functions (PDFs) $f_{Y_{i1}|\boldsymbol{B}_i}(y_{i1}|\boldsymbol{b}_i), \cdots, f_{Y_{iT_i}|\boldsymbol{B}_i}(y_{iT_i}|\boldsymbol{b}_i)$ of $Y_{i1}, \cdots, Y_{iT_i}$, given $\boldsymbol{B}_i = \boldsymbol{b}_i$, the conditional joint model of $Y_{i1}, \cdots, Y_{iT_i}$, as defined by their conditional joint PDF $f_{Y_{i1},\cdots,Y_{iT_i}|\boldsymbol{B}_i}(y_{i1}, \cdots, y_{iT_i}|\boldsymbol{b}_i) = f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i)$, with $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT_i})^\top$, is constructed by assuming conditional independence of $Y_{i1}, \cdots, Y_{iT_i}$:

$$f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) \quad = \quad \prod_{t=1}^{T_i} f_{Y_{it}|\boldsymbol{B}_i}(y_{it}|\boldsymbol{b}_i), \tag{2.2}$$

so that the resulting marginal joint model given by the marginal joint PDF $f_{Y_{i1},\cdots,Y_{iT_i}}(y_{i1}, \cdots, y_{iT_i}) = f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i)$ becomes

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) \quad = \quad \int \prod_{t=1}^{T_i} f_{Y_{it}|\boldsymbol{B}_i}(y_{it}|\boldsymbol{b}_i) f_{\boldsymbol{B}_i}(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{2.3}$$

where $f_{\boldsymbol{B}_i}(\boldsymbol{b}_i)$ is the PDF of random effects $\boldsymbol{B}_i$, $i = 1, \cdots, n$. Without random slopes of time in GLMM (1.2), it follows that $cov(Y_{it}, Y_{it'}) = cov(Y_{it}, Y_{it''})$, for any t and any $t' \neq t''$, thus ignoring the marginal longitudinal (i.e., time-varying) dependence of $Y_{i1}, \cdots, Y_{iT_i}$; such a time-independent marginal dependence structure is clearly inadequate, since, for example, we expect $cov(Y_{it}, Y_{it'}) \geq cov(Y_{it}, Y_{it''})$, for $|t - t'| \leq |t - t''|$.

Except in the case of longitudinal binary or ordinal observations $Y_{i1}, \cdots, Y_{iT_i}$, for which autoregressive LMMs for some assumed underlying latent Gaussian variables become the basis of transition models in the form of autocorrelated probit GLMMs (i.e., transition models) for $Y_{i1}, \cdots, Y_{iT_i}$ (Renard et al., 2002), exhibiting conditional dependence for non-Gaussian longitudinal data in addition to that induced by the random

effects — as in decomposition (2.1) for conventional LMMs — is not straightforward. Molenberghs and Verbeke (2005) advocate using linearization as a form of data approximation à la penalized quasi-likelihood (PQL) estimation to 'reduce' the GLMMs for the non-Gaussian data to LMMs for the corresponding 'linearized' pseudo-data. While quite general and convenient, the reliance on pseudo-data renders the model opaque; in addition, the resulting estimates tend to inherit the bias from the use of PQL.

In the sequel, we adopt a conditional version of Masarotto and Varin's (2012) Gaussian copula marginal regression model (GCMRM) as a convenient generalization of conventional LMMs and GLMMs in this context that yields a straightforward approach, akin to that for conventional LMMs, for accommodating conditional dependence in non-Gaussian longitudinal data.

3 Gaussian Copula Conditional Regression Model

To construct a conditional joint model for the non-Gaussian longitudinal observations $Y_{i1}, \cdots, Y_{iT_i}$, given the random effects $\boldsymbol{B}_i$, via their conditional joint PDF $f_{Y_{i1},\cdots,Y_{iT_i}|\boldsymbol{B}_i}(y_{i1}, \cdots, y_{iT_i}|\boldsymbol{b}_i) = f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i)$, we adopt the Gaussian copula conditional regression model (GCCRM) for $Y_{i1}, \cdots, Y_{iT_i}$ defined by

$$f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) \quad = \quad \frac{\partial}{\partial \boldsymbol{y}_i} F_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i), \tag{3.1}$$

with the conditional cumulative distribution function (CDF) $F_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) = F_{Y_{i1},\cdots,Y_{iT_i}|\boldsymbol{B}_i}(y_{i1}, \cdots, y_{iT_i}|\boldsymbol{b}_i)$ modeled by the T_i-dimensional Gaussian copula given by

$$F_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) \quad = \quad \Phi_{T_i}[q_{i1}(\mathbf{b_i}), \cdots, \mathbf{q_{iT_i}}(\mathbf{b_i}); \mathbf{R_i}], \tag{3.2}$$

where $q_{it}(\boldsymbol{b}_i) = \Phi^{-1}[u_{it}(\boldsymbol{b}_i)] = \Phi^{-1}[F_{Y_{it}|\boldsymbol{B}_i}(y_{it}|\boldsymbol{b}_i)]$ is the realization of the conditional normal score $Q_{i1}(\boldsymbol{b}_i) = \Phi^{-1}[U_{it}(\boldsymbol{b}_i)] = \Phi^{-1}[F_{Y_{it}|\boldsymbol{B}_i}(Y_{it}|\boldsymbol{b}_i)] \sim N(0,1)$, for $t = 1, \cdots, T_i$, with $F_{Y_{it}|\boldsymbol{B}_i}(y_{it}|\boldsymbol{b}_i)$ the marginal conditional CDF of Y_{it}, $\Phi_{T_i}(z_{i1}, \cdots, z_{iT_i}; \boldsymbol{R}_i)$ the CDF of the T_i-dimensional standard Gaussian distribution (i.e., with standard normal margins) with correlation matrix $\boldsymbol{R}_i$, and $\Phi^{-1}(z)$ is the quantile function of the standard normal distribution $N(0,1)$. Given $\boldsymbol{B}_i = \boldsymbol{b}_i$, note that the conditional probability integral transforms (PITs) are such that $U_{it}(\boldsymbol{b}_i) \sim U(0,1)$, with corresponding conditional normal scores $Q_{it}(\boldsymbol{b}_i) \sim N(0,1)$, for $i = 1, \cdots, n$, and for $t = 1, \cdots, T_i$, provided $Y_{i1}, \cdots, Y_{iT_i}$ are continuous random variables; this suggests that the conditional normal correlation matrix $\boldsymbol{R}_i$ containing the conditional normal correlations $r_{itt'} = corr[Q_{it}(\boldsymbol{b}_i), Q_{it'}(\boldsymbol{b}_i)]$, provides a margin-free measure of conditional dependence of $Y_{i1}, \cdots, Y_{iT_i}$. We assume $r_{itt'} = r_{i'tt'} = r_{tt'}$, for all $i \neq i'$ and for all $t, t' \leq \min(T_i, T_{i'})$, and we index $\boldsymbol{R}_i$ by i for precisely the same reason as we did $\boldsymbol{\Sigma}_i$ in (2.1); hence, we can drop the index i from $r_{itt'}$.

If Y_{it} is amenable to be modeled by a (non-Gaussian) LMM, its marginal conditional CDF $F_{Y_{it}|\boldsymbol{B}_i}(y_{it}|\boldsymbol{b}_i)$ has conditional mean given in (1.1), with possibly non-Gaussian error ϵ_{it}. For example, a logistic residual error ϵ_{it} corresponds to a marginal conditional logistic model for Y_{it}, given $\boldsymbol{B}_i$; more generally, ϵ_{it} can have a distribution from the

location-scale family, which include the Gaussian and logistic models, among others. It is easy to see that the GCCRM in (3.1) reduces to the conventional LMM for conditionally Gaussian residual error $\boldsymbol{\epsilon}_i$.

For a marginal conditional GLMM for Y_{it} with non-identity link $g(\cdot)$, its marginal conditional mean follows from (1.2) as $E(Y_{it}|\boldsymbol{B}_i) = g^{-1}(\boldsymbol{x}_{it}^\top\boldsymbol{\beta} + \boldsymbol{z}_i^\top\boldsymbol{B}_i)$. With diagonal $\boldsymbol{R}_i$, the GCCRM in (3.1) specializes into the conditional joint model (2.2) with conditionally independent marginal conditional GLMMs for $Y_{i1}, \cdots, Y_{iT_i}$.

Note that the GCCRM in (3.1) makes use of the Gaussian copula to construct the conditional joint distribution of the longitudinal observations on a single outcome from a given subject; this contrasts with Wu and de Leon's (2014) GCMM, which relies on the Gaussian copula to model the conditional dependence between multiple outcomes (but not between observations on the same outcome) in correlated data settings. Indeed, (2.3) is more akin to Masarotto and Varin's (2012) GCMRM, albeit rendered conditional by conditioning it on $\boldsymbol{B}_i$. As such, it preserves the properties of the marginal conditional LMMs/GLMMs, including the usual marginal and conditional interpretations of the model parameters.

3.1 Conditional and marginal dependence

The matrix $\boldsymbol{R}_i$ of conditional normal correlations plays the role of the residual error covariance matrix $\boldsymbol{\Sigma}_i$ in the conventional LMMs. As such, it models the conditional dependence of the non-Gaussian observations $Y_{i1}, \cdots, Y_{iT_i}$ via the conditional normal correlations $r_{tt'}$, for $t < t'$. Since $\rho_M(Y_{it}, Y_{it'}|\boldsymbol{b}_i) = |r_{tt'}|$ (Klaassen and Wellner, 1997), where $\rho_M(Y_{it}, Y_{it'}|\boldsymbol{b}_i) = \sup_{\forall a_1, a_2} corr(a_1(Y_{it}), a_2(Y_{it'})|\boldsymbol{b}_i)$ is the maximum conditional correlation between Y_{it} and $Y_{it'}$, where $a_1(y)$ and $a_2(y)$ are any functions such that $var(a_1(Y)|\boldsymbol{b}_i) < +\infty$ and $var(a_2(Y)|\boldsymbol{b}_i) < +\infty$, it follows that

$$\rho_{tt'} \;=\; corr(Y_{it}, Y_{it'}|\boldsymbol{b}_i) \;\leq\; |r_{tt'}|, \tag{3.3}$$

with $\rho_{tt'}$ the conditional correlation between Y_{it} and $Y_{it'}$. Observe that the GCCRM does not model $\rho_{tt'}$ directly and uses the normal correlation $r_{tt'}$ instead as the conditional dependence measure. Although interest is usually on $\rho_{tt'}$, $r_{tt'}$ is frequently adopted as a proxy for $\rho_{tt'}$, or as a bound for $\rho_{tt'}$, using (3.3). Alternatively, $r_{tt'}$ can be calculated from $\rho_{tt'}$ via Kugiumtzis and Bora-Senta's (2010) piecewise linear approximation method.

To mimic the AR(1) specification, for example, of $\boldsymbol{\Sigma}_i$ in conventional LMMs, we can specify a similar AR(1) specification of $\boldsymbol{R}_i = \boldsymbol{R}_i(\rho)$, where $r_{tt'} = \rho^{|t-t'|}$, for $t \neq t'$, for some $|\rho| < 1$; it is also possible to adopt a non-longitudinal specification for $\boldsymbol{R}_i$, as in the compound symmetric structure $\boldsymbol{R}_i = \rho\boldsymbol{I}_{T_i}$ (i.e., $r_{tt'} = \rho$, for all t, t').

Because $\boldsymbol{R}_i$ is the correlation matrix of the conditional normal scores — a non-linear transformation of $Y_{i1}, \cdots, Y_{iT_i}$ — an AR(1) structure, say, for $\boldsymbol{R}_i$ does not necessarily translate into an AR(1) specification for the correlation matrix of $Y_{i1}, \cdots, Y_{iT_i}$. Clemen and Reilly (1999) recommend instead the non-parametric rank-order correlations Spearman's rho ρ_S and Kendall's tau τ, which are invariant to monotonic transformations, including the conditional normal scores $Q_{it}(\boldsymbol{b}_i)$, and then using the well-known rela-

tionships between the Pearson's correlation coefficient and ρ_S and τ for the Gaussian distribution to assess $\boldsymbol{R}_i$.

The corresponding marginal longitudinal correlation $corr(Y_{it}, Y_{it'})$ is obtained via the decomposition

$$corr(Y_{it}, Y_{it'}) = \frac{cov(Y_{it}, Y_{it'})}{\sqrt{var(Y_{it})var(Y_{it'})}} = \frac{E(cov(Y_{it}, Y_{it'}|\boldsymbol{B}_i)) + cov(E(Y_{it}|\boldsymbol{B}_i), E(Y_{it'}|\boldsymbol{B}_i))}{\sqrt{var(Y_{it})var(Y_{it'})}}, \tag{3.4}$$

where $E(Y_{it}|\boldsymbol{B}_i) = \boldsymbol{x}_{it}^\top\boldsymbol{\beta} + \boldsymbol{z}_{it}^\top\boldsymbol{B}_i$ (i.e., $E(Y_{it}|\boldsymbol{B}_i)$ includes random slopes of time) in the case of a LMM for Y_{it} or $E(Y_{it}|\boldsymbol{B}_i) = g^{-1}(\boldsymbol{x}_{it}^\top\boldsymbol{\beta} + \boldsymbol{z}_{it}^\top\boldsymbol{B}_i)$ for a GLMM for Y_{it}. Although $E(Y_{it}|\boldsymbol{B}_i)$ and $cov(Y_{it}, Y_{it'}|\boldsymbol{B}_i)$ have closed-forms, getting the marginal expectation and covariance in (3.4) may be analytically intractable. One notable exception would be for the identity link (i.e., LMMs), in which case, we have

$$corr(Y_{it}, Y_{it'}) = \frac{\rho_{tt'}\sigma_t\sigma_{t'} + \boldsymbol{z}_{it}^\top\boldsymbol{\Sigma}_B\boldsymbol{z}_{it'}}{\sqrt{(\sigma_t^2 + \boldsymbol{z}_{it}^\top\boldsymbol{\Sigma}_B\boldsymbol{z}_{it})(\sigma_{t'}^2 + \boldsymbol{z}_{it'}^\top\boldsymbol{\Sigma}_B\boldsymbol{z}_{it'})}}, \quad t \le t', \tag{3.5}$$

where $\sigma_t^2 = var(Y_{it}|\boldsymbol{B}_i)$; in practice, we may simplify the model by assuming homogeneity over time (i.e., $\sigma_t^2 = \sigma^2$, for all t). The value (say, an estimate) of $\rho_{tt'}$ may be obtained from that of $r_{tt'}$ via piecewise linear approximation (Kugiumtzis and Bora-Senta, 2010).

For non-identity links (i.e., GLMMs), a crude first-order Taylor series approximation around $\boldsymbol{B}_i = \boldsymbol{0}$ can be used to approximate $corr(Y_{it}, Y_{it'})$; see, e.g., Vangeneugden et al. (2011) for details.

3.2 Likelihood estimation

Given longitudinal data $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ from the n subjects, the marginal likelihood function is then

$$L(\boldsymbol{\Theta}; \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) = \prod_{i=1}^{n} \int f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) f_{\boldsymbol{B}_i}(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{3.6}$$

where $\boldsymbol{\Theta}$ contains all the unknown parameters. We then maximize (3.6) to obtain the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$. The evaluation and maximization of (3.6) is carried out numerically using Newton-type algorithm for the case of Gaussian data, and Gaussian quadrature or importance sampling, among other methods, for the case of non-Gaussian data. We implemented this via the R function `optim` (R, 2020). To calculate the standard errors (SEs), we rely on standard likelihood theory, which states that, for large n, $\widehat{\boldsymbol{\Theta}}$ is asymptotically Gaussian with mean $\boldsymbol{\Theta}$ and covariance matrix

$$cov(\widehat{\boldsymbol{\Theta}}) = \left(E\left\{\boldsymbol{S}(\boldsymbol{\Theta})\boldsymbol{S}^\top(\boldsymbol{\Theta})\right\}\right)^{-1} = \left(E\left\{\frac{\partial}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top}\log L(\boldsymbol{\Theta}; \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)\right\}\right)^{-1}, \tag{3.7}$$

where $\boldsymbol{S}(\boldsymbol{\Theta}) = \partial \log L(\boldsymbol{\Theta}; \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)/\partial\boldsymbol{\Theta}$; the SEs of the MLEs are then calculated by evaluating (3.7) at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$ to obtain the estimated covariance matrix $\widehat{cov}(\widehat{\boldsymbol{\Theta}})$. One can also use the estimated Hessian matrix from `optim` to obtain the SEs. The R packages `lme` and `nlme`, among many others, conveniently implement these for conventional LMMs and GLMMs.

4 Simulation Study

In this section, we investigate empirically via simulations how mis-specifying the conditional dependence in longitudinal data analysis impacts estimation. We report three such simulation studies: the first involves a conventional LMM (for which conditional dependence can be easily incorporated) while the second and third concern GLMMs for non-Gaussian binary and gamma distributed outcomes, respectively (which conventionally assume conditional independence).

4.1 Simulation study 1

This study involves the conventional Gaussian LMM, for which the residual errors are conditionally, given subject-specific random effects, dependent (i.e., $\mathbf{\Sigma}_i$ is non-diagonal, $\forall i$). We compare the MLEs from the correct model that accounts for conditional dependence with those obtained "naively" based on the incorrect assumption of conditional independence. Our aim is to show the efficiency lost from incorrectly assuming conditional independence of longitudinal observations for subjects. We consider the following LMM for longitudinal observation Y_{it} at time $t(= 1, \cdots, T_i)$ from subject $i(= 1, \cdots, n)$:

$$Y_{it} \quad = \quad \beta_0 + \beta_1 \times \text{time}_{it} + B_i + \epsilon_{it}, \tag{4.1}$$

where $\text{time}_{it} = (t-1)/10$, $B_i \overset{iid}{\sim} N(0, \sigma_B^2)$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{iT_i})^\top \sim N_{T_i}(\mathbf{0}, \mathbf{\Sigma}_i)$, with $\mathbf{\Sigma}_i$ having an AR(1) structure. That is, we assume that $cov(\epsilon_{it}, \epsilon_{it'}) = cov(Y_{it}, Y_{it'}|B_i) = \sigma^2\rho^{|t-t'|}$, for $t, t' = 1, \cdots, T_i$, where $\sigma^2 = var(\epsilon_{it}) = var(Y_{it}|B_i)$ and $\rho = corr(\epsilon_{it}, \epsilon_{i,t+1}) = corr(Y_{it}, Y_{i,t+1}|B_i)$, $\forall i, t$. Note that (4.1) only includes a random intercept and no random slope of time since the conditional dependence structure is already longitudinal (i.e., $\mathbf{\Sigma}_i$ is AR(1)), which render the marginal longitudinal correlations to be time-dependent. Observe as well that (4.1) is a special case of GCCRM with Gaussian LMM margins.

In the simulations, we set $n = 200$, $T_1 = \cdots = T_n = 5$ (i.e., balanced design), $\beta_0 = 1$, $\beta_1 = 0.5$, $\sigma_B^2 = 1$, $\rho = 0.3, 0.6$, and $\sigma^2 = 0.25, 0.5, 1$. For each of $2 \times 3 = 6$ parameter settings, we generated $R = 5000$ independent datasets $\{\boldsymbol{y}_i^{(r)} = (y_{i1}^{(r)}, \cdots, y_{iT_i}^{(r)})^\top, i = 1, \cdots, n; t = 1, \cdots, T_i\}$, where $y_{it}^{(r)}$ is generated as $y_{it}^{(r)} = \beta_0 + \beta_1 \times \text{time}_{it} + B_i^{(r)} + \epsilon_{it}^{(r)}$, for $i = 1, \cdots, n$, $t = 1, \cdots, T_i$, and $r = 1, \cdots, R$. The MLE $\widehat{\mathbf{\Theta}}_c^{(r)} = (\widehat{\beta}_{c,0}^{(r)}, \widehat{\beta}_{c,1}^{(r)}, \widehat{\sigma}_{c,B}^{2(r)}, \widehat{\sigma}_c^{2(r)}, \widehat{\rho}_c^{(r)})^\top$ of $\mathbf{\Theta}_c = (\beta_0, \beta_1, \sigma_B^2, \sigma^2, \rho)^\top$, the parameter vector for the correct model, was obtained for each simulation repeat $r = 1, \cdots, R$; similarly, we also calculated the 'naive' MLE $\widehat{\mathbf{\Theta}}_m^{(r)} = (\widehat{\beta}_{m,0}^{(r)}, \widehat{\beta}_{m,1}^{(r)}, \widehat{\sigma}_{m,B}^{2(r)}, \widehat{\sigma}_m^{2(r)})^\top$ of $\mathbf{\Theta}_m = (\beta_0, \beta_1, \sigma_B^2, \sigma^2)^\top$, the parameter vector for the misspecified model that incorrectly assumes conditional independence (i.e., $\rho = 0$, when $\rho > 0$).

The bias of parameter estimates, along with their SEs, and their corresponding mean squared error (MSE) defined as MSE=bias^2 + SE^2 were calculated to evaluate the bias and relative efficiency of the MLEs. The results, obtained using the R package `lme`, are reported in Table 1 for different values of σ^2 and ρ.

Table 1: Bias estimates, with their standard errors (SEs), and corresponding mean squared error (MSE), based on $R = 5000$ simulated datasets from LMM (4.1).

Parameter			Bias		SE		MSE	
			Mis-specified	Correct	Mis-specified	Correct	Mis-specified	Correct
$\sigma^2 = 0.25$	$\rho = 0.3$	$\beta_0 = 1$	-0.0004	-0.0004	0.0770	0.0767	0.006	0.006
		$\beta_1 = 0.5$	0.0008	0.0007	0.1222	0.1211	0.015	0.015
		$\sigma^2_B = 1$	0.0320	-0.0065	0.1094	0.1101	0.013	0.012
		$\sigma^2 = 0.25$	-0.0385	0.0004	0.0110	0.0189	0.002	0.0004
		$\rho = 0.3$	—	-0.0023	—	0.0538	—	0.003
	$\rho = 0.6$	$\beta_0 = 1$	-0.0004	-0.0004	0.0785	0.0779	0.006	0.006
		$\beta_1 = 0.5$	0.0007	0.0007	0.1190	0.1164	0.014	0.014
		$\sigma^2_B = 1$	0.0950	-0.0094	0.1145	0.1185	0.022	0.014
		$\sigma^2 = 0.25$	-0.0013	0.0027	0.0086	0.0372	0.0001	0.001
		$\rho = 0.6$	—	-0.0036	—	0.0575	—	0.003
$\sigma^2 = 0.5$	$\rho = 0.3$	$\beta_0 = 1$	-0.0005	-0.0005	0.0833	0.0829	0.007	0.007
		$\beta_1 = 0.5$	0.0011	0.0010	0.1728	0.1712	0.030	0.029
		$\sigma^2_B = 1$	0.0700	-0.0077	0.1178	0.1203	0.019	0.020
		$\sigma^2 = 0.5$	-0.0770	0.0008	0.0220	0.0377	0.006	0.014
		$\rho = 0.3$	—	-0.0022	—	0.0536	—	0.003
	$\rho = 0.6$	$\beta_0 = 1$	-0.0005	-0.0005	0.0860	0.0851	0.007	0.007
		$\beta_1 = 0.5$	0.0010	0.0010	0.1682	0.1647	0.028	0.027
		$\sigma^2_B = 1$	0.1950	-0.0133	0.1279	0.1417	0.054	0.020
		$\sigma^2 = 0.5$	-0.0026	0.0054	0.0171	0.0737	0.0003	0.005
		$\rho = 0.6$	—	-0.0035	—	0.0570	—	0.003
$\sigma^2 = 1$	$\rho = 0.3$	$\beta_0 = 1$	-0.0006	-0.0006	0.0947	0.0940	0.009	0.009
		$\beta_1 = 0.5$	0.0015	0.0015	0.2444	0.2422	0.060	0.059
		$\sigma^2_B = 1$	0.1460	-0.0100	0.1344	0.1426	0.039	0.020
		$\sigma^2 = 1$	-0.1540	0.0020	0.0439	0.0752	0.026	0.006
		$\rho = 0.3$	—	-0.0022	—	0.0532	—	0.003
	$\rho = 0.6$	$\beta_0 = 1$	-0.0008	-0.0008	0.1004	0.0988	0.010	0.010
		$\beta_1 = 0.5$	0.0018	0.0016	0.2381	0.2330	0.057	0.054
		$\sigma^2_B = 1$	0.3950	-0.0211	0.1558	0.1964	0.180	0.039
		$\sigma^2 = 1$	-0.4054	0.0100	0.0352	0.1462	0.166	0.021
		$\rho = 0.6$	—	-0.0036	—	0.0570	—	0.003

As expected, the 'naive' MLEs based on the mis-specified model generally exhibited more bias and less efficiency than those based on the correct model. In particular, we observed that regardless of the magnitude of σ^2, the 'naive' MLEs of variance components σ_B^2 and σ^2 had non-negligible bias and lower efficiency (i.e., SEs of 'naive' MLEs were generally smaller than those for the 'correct' MLEs), the latter a common consequence of an incorrect independence assumption on the data. In particular, magnitude of the corresponding MSE of $\hat{\sigma}_B^2$ and $\hat{\sigma}^2$ in the case of large σ^2 and ρ are much bigger for the 'naive' method compared to the GCCRM approach. We should note, however, that the regression coefficients were estimated generally well by both 'naive' and 'correct' MLEs; being marginal parameters, this is to be expected. However, while the 'naive' estimates of the regression coefficients remained unbiased, the corresponding SEs generally tended to be smaller than what they should be, when $\rho > 0$.

Our results clearly show that if one ignores the conditional dependence between longitudinal observations from subjects, one runs the risk of ending up with incorrect conclusions; this is true even if interest lies only on the regression coefficients, since the 'naive' SEs tend to underestimate the true sampling variability of the estimates, which in turn can lead to tests with inflated Type I error rates or CIs that are misleadingly narrow.

4.2 Simulation study 2

We also conduct a simulation study to evaluate performance of the MLE of model parameters for the GCCRM with logistic (conditional) margins. In particular, the finite-sample performance of the 'naive' MLEs, based on the mis-specified model that incorrectly assumes conditional independence, is compared with that of the MLEs based on the correct model (GCCRM). Specifically, we consider the following conditional marginal logistic model for longitudinal observation Y_{it} at time $t(= 1, \cdots, T_i)$ from subject $i(= 1, \cdots, n)$:

$$Y_{it}|B_i \sim \text{logistic}\left[\mu_{it}(B_i), \zeta\right], \tag{4.2}$$

where μ_{it} and ζ are location and scale of the logistic distribution with $\mu_{it}(B_i) = \beta_0 + \beta_1 x_{it} + B_i, B_i \overset{iid}{\sim} N(0, \sigma_B^2)$, and β_0, β_1 are intercept and slope of covariate $x_{it}(= t/10)$. We adopt a GCCRM with conditional logistic margins given by (4.2) and with $\boldsymbol{R}_i$ having an AR(1) to incorporate conditional dependence among the longitudinal observations $Y_{i1}, \cdots, Y_{iT_i}$. Since no random slope of time is included in (4.2), such an $\boldsymbol{R}_i$ guarantees that the marginal normal correlations $r_{tt'}$ (hence, the marginal longitudinal correlations $\rho_{tt'}$) are time-dependent. Note that the conventional logistic model specification of conditional independence for $Y_{i1}, \cdots, Y_{iT_i}$ corresponds to a diagonal $\boldsymbol{R}_i$.

With $\beta_0 = 1, \beta_1 = 2, \zeta = 0.1$, and four scenarios for $\sigma_B^2 = 0.1, 0.5, 0.7, 1$ and two scenarios for $\rho = corr(Q_{it}(b_i), Q_{i,t+1}(b_i)) = 0.5, 0.9$, $\forall i, t$, we generate $R = 1000$ datasets each with $n = 200$ subjects with five follow-ups per subject (balanced design) for each of $4 \times 2 = 8$ parameter settings. For each generated dataset, parameters are estimated for the both naive and GCCRM using the R package `copula` to implement the simulations. Tables 2 and 3 show the bias estimates along with their SEs and corresponding MSEs

Table 2: Bias estimates, with their standard errors (SEs), and corresponding mean squared error (MSE), based on $R = 1000$ simulated datasets from the GCCRM with conditional logistic margins (4.2) in the case of $\rho = 0.5$.

	Parameter	Bias		SE		MSE	
		Mis-specified	Correct	Mis-specified	Correct	Mis-specified	Correct
$\sigma_B^2 = 0.1$	$\beta_0 = 1$	–0.002	0.004	0.072	0.062	0.005	0.004
	$\beta_1 = 2$	–0.001	–0.016	0.146	0.113	0.021	0.013
	$\zeta = 0.1$	–0.002	0.000	0.018	0.017	0.0003	0.0003
	$\sigma_B^2 = 0.1$	–0.002	–0.001	0.019	0.018	0.0004	0.0003
	$\rho = 0.5$	—	0.000	—	0.034	—	0.001
$\sigma_B^2 = 0.5$	$\beta_0 = 1$	–0.013	0.006	0.241	0.172	0.058	0.030
	$\beta_1 = 2$	0.021	–0.038	0.487	0.287	0.238	0.084
	$\zeta = 0.1$	0.003	0.007	0.040	0.042	0.002	0.002
	$\sigma_B^2 = 0.5$	–0.031	–0.018	0.142	0.100	0.021	0.010
	$\rho = 0.5$	—	–0.002	—	0.091	—	0.008
$\sigma_B^2 = 0.7$	$\beta_0 = 1$	–0.010	0.014	0.289	0.213	0.084	0.046
	$\beta_1 = 2$	0.021	–0.057	0.583	0.361	0.340	0.134
	$\zeta = 0.1$	0.006	0.010	0.055	0.052	0.003	0.003
	$\sigma_B^2 = 0.7$	–0.068	–0.034	0.204	0.160	0.046	0.027
	$\rho = 0.5$	—	–0.004	—	0.107	—	0.011
$\sigma_B^2 = 1$	$\beta_0 = 1$	0.002	0.000	0.308	0.225	0.095	0.051
	$\beta_1 = 2$	0.024	–0.063	0.628	0.368	0.395	0.139
	$\zeta = 0.1$	0.012	0.019	0.066	0.068	0.004	0.005
	$\sigma_B^2 = 1$	–0.092	–0.043	0.299	0.223	0.098	0.052
	$\rho = 0.5$	—	–0.014	—	0.113	—	0.013

in the case of $\rho = 0.5$ and 0.9, respectively. Based on the results, it appears that the ML estimates for the GCCRM perform well unlike the 'naive' estimates in terms of MSE for the all model parameters. Our results clearly suggest the importance of accounting for conditional dependence among longitudinal observations from subjects. One may lead to wrong conclusions by ignoring conditional dependence by having severe bias in estimates and/or incorrect SEs.

4.3 Simulation study 3

We also conduct a simulation study using a non-Gaussian gamma-distributed outcome to compare the finite-sample performance of the 'naive' MLEs, based on the mis-specified model that incorrectly assumes conditional independence, with that of the MLEs based on the correct model. Specifically, we consider the following conditional marginal GLMM for longitudinal observation Y_{it} at time $t(= 1, \cdots, T_i)$ from subject $i(= 1, \cdots, n)$:

$$Y_{it}|B_i \quad \sim \quad \text{gamma}\Big[\alpha, \frac{1}{\alpha}\mu_{it}(B_i)\Big], \tag{4.3}$$

Table 3: Bias estimates, with their standard errors (SEs), and corresponding mean squared error (MSE), based on $R = 1000$ simulated datasets from the GCCRM with conditional logistic margins (4.2) in the case of $\rho = 0.9$.

	Parameter	Bias		SE		MSE	
		Mis-specified	Correct	Mis-specified	Correct	Mis-specified	Correct
$\sigma_B^2 = 0.1$	$\beta_0 = 1$	0.001	0.005	0.074	0.075	0.005	0.006
	$\beta_1 = 2$	–0.005	–0.017	0.025	0.020	0.025	0.020
	$\zeta = 0.1$	–0.002	–0.002	0.020	0.017	0.0004	0.0003
	$\sigma_B^2 = 0.1$	–0.004	–0.002	0.022	0.018	0.0005	0.0003
	$\rho = 0.9$	—-	–0.013	—-	0.062	—-	0.004
$\sigma_B^2 = 0.5$	$\beta_0 = 1$	0.002	–0.002	0.246	0.163	0.060	0.026
	$\beta_1 = 2$	–0.005	–0.026	0.502	0.256	0.252	0.066
	$\zeta = 0.1$	0.006	0.009	0.046	0.040	0.002	0.002
	$\sigma_B^2 = 0.5$	–0.034	–0.008	0.128	0.109	0.018	0.012
	$\rho = 0.9$	—-	–0.032	—-	0.123	—-	0.016
$\sigma_B^2 = 0.7$	$\beta_0 = 1$	–0.005	0.012	0.247	0.197	0.061	0.039
	$\beta_1 = 2$	–0.001	–0.052	0.502	0.327	0.252	0.110
	$\zeta = 0.1$	0.009	0.012	0.054	0.052	0.003	0.003
	$\sigma_B^2 = 0.7$	–0.046	–0.024	0.182	0.150	0.035	0.023
	$\rho = 0.9$	—-	–0.040	—-	0.144	—-	0.022
$\sigma_B^2 = 1$	$\beta_0 = 1$	0.010	0.001	0.288	0.213	0.083	0.045
	$\beta_1 = 2$	–0.011	–0.046	0.589	0.354	0.347	0.127
	$\zeta = 0.1$	0.011	0.018	0.066	0.068	0.004	0.005
	$\sigma_B^2 = 1$	–0.081	–0.045	0.289	0.221	0.090	0.051
	$\rho = 0.9$	—-	–0.046	—-	0.152	—-	0.025

with

$$\mu_{it}(B_i) \quad = \quad E(Y_{it}|B_i) \quad = \quad \exp(\beta_0 + \beta_1 x_{it} + B_i), \tag{4.4}$$

where α is the shape parameter and $B_i \overset{iid}{\sim} N(0, \sigma_B^2)$, with covariate x_{it} generated from $U(0, 2)$. To incorporate conditional dependence among the longitudinal observations $Y_{i1}, \cdots, Y_{iT_i}$, we adopt a GCCRM with GLMM margins given by (4.3) and with $\boldsymbol{R}_i$ having an AR(1) structure. Since no random slope of time is included in (4.3), such an $\boldsymbol{R}_i$ guarantees that the marginal normal correlations $r_{tt'}$ (hence, the marginal longitudinal correlations $\rho_{tt'}$) are time-dependent. Note that the conventional GLMM specification of conditional independence for $Y_{i1}, \cdots, Y_{iT_i}$ corresponds to a diagonal $\boldsymbol{R}_i$.

We independently generate $R = 1000$ datasets with $n = 200$ and $T_1 = \cdots = T_n = 5$ (i.e., balanced design) from a GCCRM with conditional margins given by (4.3) and parameters $\beta_0 = 1, \beta_1 = -1, \alpha = 2$, and $\sigma_B^2 = 0.02$. For $\boldsymbol{R}_i$, we set $\rho = corr(Q_{it}(b_i), Q_{i,t+1}(b_i)) = 0.2$, $\forall i, t$. We obtained the MLE $\widehat{\boldsymbol{\Theta}}_c^{(r)} = (\widehat{\beta}_{c,0}^{(r)}, \widehat{\beta}_{c,1}^{(r)}, \widehat{\alpha}_c^{(r)}, \widehat{\sigma}_{c,B}^{2(r)}, \widehat{\rho}_c^{(r)})^\top$ from the correct model with $\rho = 0.2$ as well as the 'naive' MLE $\widehat{\boldsymbol{\Theta}}_m^{(r)} = (\widehat{\beta}_{m,0}^{(r)}, \widehat{\beta}_{m,1}^{(r)}, \widehat{\alpha}_m^{(r)}, \widehat{\sigma}_{m,B}^{2(r)})^\top$ from the mis-specified model with $\rho = 0$, for each simulation repeat $r = 1, \cdots, R$. We used the R packages `copula` and `glmer` (also `lme4`) to implement the simulations.

Table 4: Bias estimates, with their standard errors (SEs), and corresponding mean squared error (MSE), based on $R = 1000$ simulated datasets from the GCCRM with GLLM margins (4.3).

Parameter	Bias		SE		MSE	
	Mis-specified	Correct	Mis-specified	Correct	Mis-specified	Correct
$\beta_0 = 1$	–0.680	0.000	0.049	0.025	0.465	0.001
$\beta_1 = -1$	2.003	–0.016	0.039	0.013	4.014	0.0004
$\alpha = 2$	0.333	0.081	0.118	0.013	0.125	0.007
$\sigma_B^2 = 0.02$	0.059	0.002	0.012	0.009	0.004	0.0001
$\rho = 0.2$	—	0.006	—	0.014	—	0.0002

Table 4 displays the bias estimates along with their SEs, and corresponding MSEs. Note that unlike the MLEs from the correct model, which are fairly unbiased, the 'naive' estimates suffer from severe bias with also larger MSEs. The large bias exhibited even by the 'naively' estimated fixed effects $\widehat{\beta}_{m,0}$ and $\widehat{\beta}_{m,1}$ can be explained by noting that these fixed effects are not the marginal effects because we have a non-identity link function, and hence, the marginal mean model is likewise mis-specified whenever the conditional dependence structure is mis-specified. We also observed similar behaviour for different values of σ_B^2 and ρ as in the simulation study 2.

As in simulation studies 1 and 2, our results clearly suggest the importance of accounting for conditional dependence among the longitudinal observations from subjects. Failure to do so is likely to yield severe bias in estimates and incorrect SEs, thus leading to possibly misleading conclusions.

5 Applications

5.1 Bolus count data

We now illustrate the GCCRM methodology on bolus count data, which have been analyzed by Weiss (2005). This data is a study of 65 patients ($n = 65$) controlled analgesia comparing two different dosing regimes. In particular, there are two groups, a 1 milligram (mg) per dose group and a 2 mg per dose group. After each dose, there is a lockout time where the patient may not administer more medication, noting that the lockout time in the 2 mg dose group is twice as long as in the 1 mg dose group. The number of doses is recorded for 12 consecutive ($T = 12$) 4-hour periods where the lockout time allows for a maximum of 30 dosages in the 2 mg group and 60 dosages in the 1 mg group. Note that there were no responses near the upper limit such that the maximum count in each group for each four hour period was less than the theoretical maximum number of dosages.

Following Weiss (2005), we use the Poisson distribution to model the counts. It is observed that the counts are higher for the 1 mg group as might be expected, although the amount does not seem to be twice that in the 2 mg group. It seems that the counts are decreasing over time, and there may be a bump up in counts at about $t = 5$ and $t = 10$, and in particular for the 2 mg group in most cases (see Weiss, 2005 for more details).

We fit a random intercept model to this data. Following Weiss (2005), for the fixed effects, we keep a constant difference between groups and an unstructured mean for the time trend, as the increases at times 5 and 10 are otherwise difficult to model. In particular, we consider the following model:

$$Y_{it}|B_i \sim \text{Poisson}\left[\lambda_{it}(B_i)\right], \qquad i = 1, ..., n; t = 1, ..., T, \tag{5.1}$$

where $log[\lambda_{it}(B_i)] = \alpha_t + \alpha_{13}x_{it} + B_i$, with $x_{it} = 1\{1 \text{ mg group}\}$ and $B_i \overset{iid}{\sim} N(0, \sigma_B^2)$. The parameter α_{13} is the increase from 2 mg to the 1 mg group, and $\alpha_j, (j = 1, ..., T)$, is the parameter at time j for the 2 mg group. Because the marginal GLMM (5.1) has only a random intercept, we specify an AR(1) structure for the conditional normal correlation matrix $\boldsymbol{R}_i$ of the GCCRM with GLMM conditional margins given by (5.1), for $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iT})^\top$ to capture the time-varying marginal longitudinal correlations between any longitudinal pair of observations Y_{it} and $Y_{it'}$, for $t < t'$. We implemented our analysis via the R package `copula`.

Table 5 provides the parameter estimates and corresponding standard errors based on the GCCRM and naive method (Weiss, 2005) which ignores the possible conditional dependence in the model. It is clear from Table 5 that conditional dependence among the longitudinal observations from patients need to be accounted for in the analysis as the estimated conditional normal correlation $\widehat{\rho} = \widehat{corr}(Q_{it}(b_i), Q_{i,t+1}(b_i)) = 0.56$ which is statistically significant from zero and strong enough to impact the results of the analyses based on GCCRM and naive method which assumes conditional independence. In particular, the parameter α_{13} is different based on GCCRM and naive method in terms of magnitude and sign. Also, there are differences in SEs of the estimates.

We can conclude that in the absence of a random slope of time in the marginal GLMM for Y_{it}, a naive analysis that ignores the conditional dependence among the longitudinal data yields marginal longitudinal correlations that are not time-dependence. Consequently, a GCCRM with AR(1) specification for $\boldsymbol{R}_i$, $\forall i$, avoids this and allows for a time-varying marginal longitudinal correlation structure.

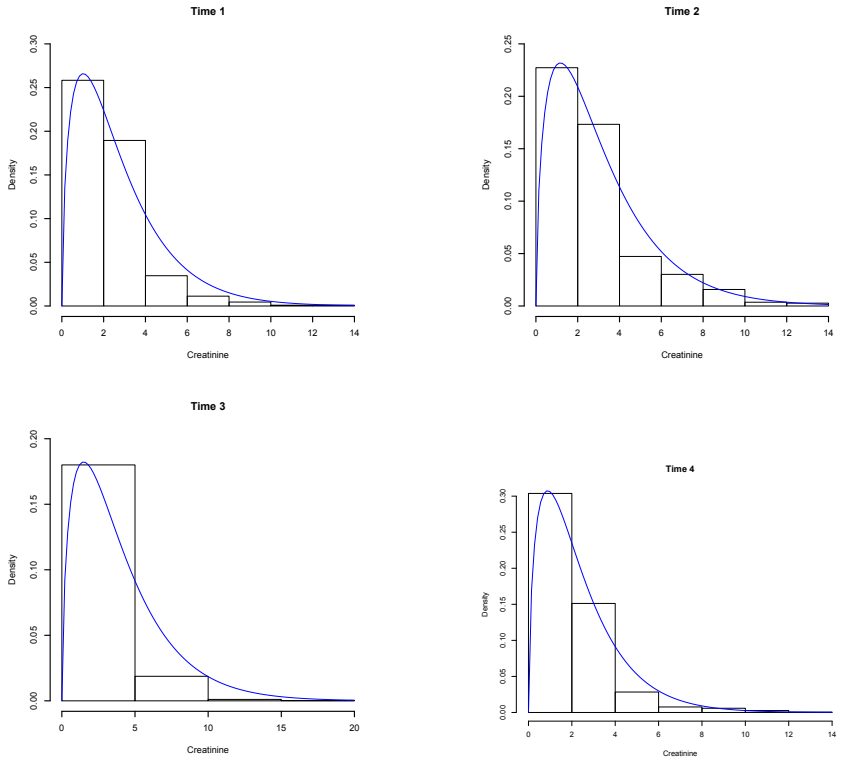
5.2 Serum creatinine data

We next illustrate the GCCRM methodology on data on serum creatinine, which are a part of a larger dataset previously analyzed by Fieuws and Verbeke (2008). To predict renal graft failure, Fieuws and Verbeke (2008) adopted a multivariate longitudinal model using markers such as serum creatinine, urine proteinuria, mean of systolic and diastolic blood pressure, and blood hematocrit level as predictors (see Fieuws and Verbeke, 2008 for more details). Fieuws and Verbeke (2008) considered patients who

Table 5: Longitudinal analysis of bolus count data based on a GCCRM with conditional marginal GLMMs given by (5.1). Also included is the 'naive' analysis based on the assumption of conditional independence of the longitudinal data.

Parameter	GCCRM		Naive	
	Estimate	SE	Estimate	SE
α_1	2.09	0.011	2.00	0.054
α_2	1.32	0.010	1.51	0.011
α_3	1.34	0.012	1.63	0.055
α_4	1.91	0.040	1.72	0.034
α_5	1.98	0.052	1.88	0.059
α_6	2.20	0.026	1.58	0.124
α_7	1.18	0.035	1.40	0.019
α_8	0.98	0.032	1.29	0.076
α_9	1.49	0.005	1.28	0.038
α_{10}	1.47	0.009	1.46	0.066
α_{11}	1.36	0.012	1.32	0.104
α_{12}	1.46	0.024	1.28	0.084
α_{13}	–0.13	0.012	0.27	0.087
σ_B^2	0.40	0.039	0.25	0.064
ρ	0.56	0.016	—	—

Figure 1: Histograms for serum creatinine with superimposed gamma density for different time periods.



received a kidney transplant and who were intensively monitored during the years after the transplant. These patients underwent, between January 21, 1983, and August 16, 2000, a primary renal transplantation with a graft from a deceased or living donor in the University Hospital Gasthuisberg at the Catholic University of Leuven in Belgium.

We considered serum creatinine as the outcome, measured on $n = 1111$ subjects with different follow-up times over a 17-year period (i.e., 1983–2000). For convenience, we divided the follow-up times into 4 based on quartiles of the follow-up times and calculated the average of the observations (including the corresponding covariates) in each of the $T_1 = \cdots = T_n = 4$ new revised follow-up times. Histograms of serum creatinine at the revised follow-up times, shown in Figure 1, suggest the distribution of serum creatinine is right-skewed.

Let Y_{it} be the serum creatinine of patient $i(= 1, \cdots, n)$ at follow-up time $t(= 1, \cdots, T_i)$. We assume the same conditional marginal GLMM for Y_{it} as in (4.3):

$$Y_{it}|B_i \sim \text{gamma}\Big[\alpha, \frac{1}{\alpha}\mu_{it}(B_i)\Big], \tag{5.2}$$

Table 6: Longitudinal analysis of serum creatinine data based on a GCCRM with conditional marginal GLMMs given by (5.2). Also included is the 'naive' analysis based on the assumption of conditional independence of the longitudinal data.

Parameter	GCCRM		Naive	
	Estimate	SE	Estimate	SE
constant	0.5000	0.0130	0.4770	0.0440
sex	–0.0790	0.0300	–0.0570	0.0120
BMI	0.0100	0.0160	0.0030	0.0010
SBP	–0.0001	0.0050	–0.0002	0.0002
α	1.4980	0.0930	3.8500	1.9610
σ_B^2	0.0070	0.0120	0.0170	0.1300
ρ	0.5000	0.0250	—	—

where $\mu_{it}(B_i) = \exp(\boldsymbol{x}_{it}^\top \boldsymbol{\beta} + B_i)$, with $B_i \overset{iid}{\sim} N(0, \sigma_B^2)$. The vector $\boldsymbol{x}_{it}$ of covariates consists of sex (sex_i), body mass index (BMI_{it}), and systolic blood pressure (SBP_{it}), which have all been shown to be excellent predictors of a patient's serum creatinine (Young, 2002). Because the marginal GLMM (5.2) has only a random intercept, we specify an AR(1) structure for the conditional normal correlation matrix $\boldsymbol{R}_i$ of the GCCRM with GLMM conditional margins given by (5.2), for $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iT_i})^\top$ to capture the time-varying marginal longitudinal correlations between any longitudinal pair of observations Y_{it} and $Y_{it'}$, for $t < t'$. We implemented our analysis via the R packages `copula`, `lme4`, and `glmer`. Results are shown in Table 6.

It is clear from Table 6 that conditional dependence among the longitudinal observations from subjects needs to be accounted for in the analysis; the estimated conditional normal correlation $\widehat{\rho} = \widehat{corr}(Q_{it}(b_i), Q_{i,t+1}(b_i)) = 0.5$ is statistically significant and strong enough to impact the results of the analyses based on GCCRM and on the 'naive' assumption of conditional independence. This can be seen in the differences in SEs of the estimates: because GCCRM accounts for conditional dependence, thus adding another source of variation to the analysis, the 'naive' analysis yielded possibly deflated SEs — relative to those from GCCRM — for the regression coefficients corresponding to the covariates sex, BMI and SBP. Computation of the exact marginal correlations in this case is tedious; hence, one can use the first-order Taylor series-based approximation to show that $corr(Y_{it}, Y_{it'}) \approx \frac{\alpha\sigma_B^2}{\alpha\sigma_B^2+1}$, $t \leq t'$, which is 0.010 in our data analysis.

In the absence of a random slope of time in the marginal GLMM for Y_{it}, a 'naive' analysis that ignores the conditional dependence among the longitudinal data yields marginal longitudinal correlations that are not time-dependent; that is, it assumes $corr(Y_{it}, Y_{it'}) = corr(Y_{it}, Y_{it''}) \approx 0.061$, $\forall i, t, t', t''$, an obviously inadequate model of the correlation structure for temporally-ordered data. A GCCRM with AR(1) specification for $\boldsymbol{R}_i$, $\forall i$, avoids this and allows for a time-varying marginal longitudinal correlation

structure.

6 Conclusion

Mixed models are widely used to analyze longitudinal data, especially in health and medical research. In their conventional formulation as LMMs and GLMMs, a commonly indispensable assumption in settings involving longitudinal non-Gaussian data is that the longitudinal observations from subjects are conditionally independent, given subject-specific random effects. Although conventional Gaussian LMMs are able to easily incorporate conditional dependence of longitudinal observations, they require that the data are, or some transformation of them is, Gaussian, an admittedly serious limitation in a wide variety of practical applications.

In this paper, we introduced the class of GCCRMs as flexible alternatives to conventional LMMs and GLMMs. The proposed GCCRM is different from the GCMRM (Masarotto and Varin, 2012) and GCMM (Wu and de Leon, 2014). First, unlike GCMRMs, which model population-averaged effects of covariates, GCCRMs are subject-specific models that account for subject-specific heterogeneity; indeed, GCCRMs are GCMRMs rendered conditional by conditioning GCMRMs on subject-specific random effects. Second, GCCRMs consider only a single non-Gaussian outcome while Wu and de Leon's (2014) GCMMs involve multiple (e.g., mixed binary and continuous) non-Gaussian outcomes in a cluster (e.g., longitudinal) setting. As such, GCCRMs need only to account for conditional dependence of the longitudinal observations via the Gaussian copula; GCMMs, on the other hand, incorporate conditional dependence between only the outcomes via the Gaussian copula but not among the repeat (i.e., longitudinal) observations, which are still assumed to be conditionally independent.

One particularly attractive property of GCCRMs is that they extend conventional LMMs and GLMMs in a conveniently natural way that reduces to conventional LMMs, when the data are Gaussian, and to conventional GLMMs, when conditional independence is assumed. Likelihood analysis of GCCRMs can also be implemented using existing software and statistical packages.

We evaluated the finite-sample performance of MLEs for GCCRM empirically via simulations vis-à-vis the 'naive' likelihood analysis that incorrectly assumes conditionally independent longitudinal data. Our results showed that the 'naive' analysis tends to yield estimates with possibly severe bias and incorrect SEs. Finally, we illustrated the proposed methodology on two datasets: bolus count data on patients' controlled analgesia comparing two different dosing regimes (Weiss, 2005), and another data on serum creatinine from a study on renal graft failure (Fieuws and Verbeke, 2008). We concluded that ignoring conditional dependence of longitudinal data may result in wrong conclusions and inferences from tests and confidence intervals.

Acknowledgements

This work was partially supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada. Part of it was carried out while ARdL was a Visiting Professor at the School of Statistics, University of the Philippines. MT and ARdL are grateful to Dr. Steffen Fieuws of Interuniversity Institute for Biostatistics and Statistical Bioinformatics at Katholieke Universiteit Leuven for kindly providing them the renal graft data, from which the serum creatinine data came.

Supplementary Materials

The Supplementary Materials contain R codes and corresponding “readme" files for the simulations and applications conducted in this article.

References

[1] Brown, H. and Prescott, R. (2015). *Applied Mixed Models in Medicine*. New York: John Wiley & Sons.

[2] Clemen, R. and Reilly, T. (1999). Correlation and copulas for decision and risk analysis. *Management Science* **45**, 208–224.

[3] Das, K., Li, R., Sengupta, S., and Wu, R. (2013). A Bayesian semiparametric model for bivariate sparse longitudinal data. *Statistics in Medicine* **32**, 3899–3910.

[4] de Leon, A. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* **30**, 175–185.

[5] Fieuws, S. and Verbeke, G. (2008). Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* **9**, 419–431.

[6] Klaassen, C. and Wellner, J. (1997). Efficient estimation in the bivariate normal copula: normal margins are least favourable. *Bernoulli* **3**, 55–77.

[7] Kugiumtzis, D. and Bora-Senta, E. (2010). Normal correlation coefficient of non-normal variables using piece-wise linear approximation. *Computational Statistics* **25**, 645–662.

[8] Magezi, D.A. (2015). Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology* **6:2**, doi:10.3389/fpsyg.2015.00002.

[9] Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517–1549.

[10] McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

[11] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

[12] R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

[13] Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate end points in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal* **44**, 921–935.

[14] Searle, S.R., Casella, G., and McCulloch, C.E. (2006). *Variance Components*. New Jersey: John Wiley & Sons, doi:10.1375/twin.14.1.25.

[15] Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demetrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics* **38**, 215–232.

[16] Weiss, E.W. (2005). *Modeling Longitudinal Data*. New York: Springer.

[17] Wu, B. and de Leon, A.R. (2014). Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *Journal of Agricultural, Biological and Environmental Statistics* **19**, 39–56.

[18] Wu, B., de Leon, A., and Withanage, N. (2013). Joint analysis of mixed discrete and continuous outcomes via copulas. In *Analysis of Mixed Data: Methods and Applications*, de Leon A and Carrie're Chough K (eds), 139–156, Chap 10. CRC/Chapman & Hall.

[19] Young, J. H. (2002). Blood pressure and decline in kidney function: findings from the systolic hypertension in the elderly program (SHEP). *Journal of the American Society of Nephrology* **11**, 2776–2782.