Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 1, 35–50
https://doi.org/10.1214/18-BJPS416

A joint mean-correlation modeling approach for longitudinal zero-inflated count data

Weiping Zhang, Jiangli Wang Fang Qian and Yu Chen

University of Science and Technology of China

Abstract. Longitudinal zero-inflated count data are widely encountered in many fields, while modeling the correlation between measurements for the same subject is more challenge due to the lack of suitable multivariate joint distributions. This paper studies a novel mean-correlation modeling approach for longitudinal zero-inflated regression model, solving both problems of specifying joint distribution and parsimoniously modeling correlations with no constraint. The joint distribution of zero-inflated discrete longitudinal responses is modeled by a copula model whose correlation parameters are innovatively represented in hyper-spherical coordinates. To overcome the computational intractability in maximizing the full likelihood function of the model, we further propose a computationally efficient pairwise likelihood approach. We then propose separated mean and correlation regression models to model these key quantities, such modeling approach can also handle irregularly and possibly subject-specific times points. The resulting estimators are shown to be consistent and asymptotically normal. Data example and simulations support the effectiveness of the proposed approach.

1 Introduction

Longitudinal zero-inflated count data are widely encountered in the fields of biomedical, medical, public health and social survey, etc. Zero inflation describes data for which the number of observed zeroes is higher than what is expected from a standard Poisson distribution and often results in over-dispersion. For example, Bulsara et al. (2004) reported a study of evaluating risk factors associated with severe hypoglycaemia (an event leading to loss of consciousness or seizure), prospective assessment of severe hypoglycaemia was made over 9-year period for a total of 1229 children with Type 1 diabetes. Patients were seen every 3 months and episodes of hypoglycaemia along with clinical data were recorded. Over 70% of children never experienced a severe hypoglycaemic event. With measuring the variables of interest longitudinally, it is important to properly address the problem of zero inflation and correlated observations within individuals over time, otherwise, the analysis may lead to biased estimates, underestimated standard errors and distorted test statistics of overall goodness of fit (Atkins and Gallop (2007)).

Recently, several so-called two-part mixed models have been developed for longitudinal zero-inflated data, which impose a binomial distribution to deal with zero versus nonzero, and other distribution (discrete distribution for counts data, or censored continuous distribution for continuous outcome) to deal with the nonzero part of the distribution. The random effect is then included either or both in the zero and nonzero parts to account for the correlation between measurements upon the same subject at different occasions. For example, Berk and Lachenbruch (2002) used a two-part model to handle the zero and positive repeated measures by including a random effect in the logistic regression part for zeros, and imposing a left censored lognormal distribution for the nonzero positive observations; For count

Key words and phrases. Copula, hyperspherical coordinates, mean-correlation regression, pairwise likelihood, zero inflated negative binomial.

Received October 2017; accepted August 2018.

data, Min and Agresti (2005) proposed a two-part zero-inflated random effects hurdle model (Mullahy (1986)) to handle the zero and positive count separately; Lee et al. (2006) incorporated shared subject-specific random effects in each zero-inflated Poisson (ZIP) regression model part to account for zero inflation and over-dispersion within longitudinal count measurements. Bulsara et al. (2004) found that the negative binomial or zero inflated models are more appropriate than the commonly used Poisson regression models for longitudinal severe hypoglycaemia data. Rose et al. (2006) argued that ZINB and Negative Binomial Hurdle (NBH) models produce more reliable inference, and which one should be used depend on the study design and purpose. Similar results can be found in Lewsey and Thomson (2004), Ground and Koch (2008) and Buu et al. (2012). Ghosh, Mukhopadhyay and Lu (2006) proposed a Bayesian alternative approach for zero-inflated regression models; Alfo and Maruotti (2010) discussed semiparametric estimation method for dynamic two-part models.

However, in many practical applications the assumption of normality and the condition of being uncorrelated between random effects and errors may be violated as the data often exhibit skewness and some covariates may be measured with measurement errors. An attractive alternative is to characterize the covariations for those repeated measurements using parsimonious regression techniques. For continuous longitudinal responses, Pourahmadi (1999, 2000) developed a modified Cholesky decomposition on covariances that allows unconstrained parametrization of the entries in the decomposition and permits the development of interpretable regression models. Along this line, joint mean-covariance modelling approaches have attracted increasing interest. See, for example, Pan and Mackenzie (2003), Ye and Pan (2006), Pourahmadi (2007), Leng, Zhang and Pan (2010), Zhang and Leng (2012), Liu and Zhang (2013), Liu, Zhang and Chen (2018). More recently, Zhang, Leng and Tang (2015) proposed models to investigate marginal variances and correlations from a geometric perspective. For discrete longitudinal responses, however, modeling the covariance is more challenging mainly because of the lack of suitable multivariate joint distributions that can support complex correlation structures. As a consequence, models for longitudinal count data have either sacrificed generality or have been specified with potentially undesirable correlation restrictions imposed for the sake of retaining computational tractability. For example, multivariate Poisson models are often constructed by including a common Poisson process that enters every outcome in the model, but this approach can only produce nonnegative correlations (Kocherlakota and Kocherlakota (1992), Karlis (2003)). It is also known that even for given marginal distributions of the discrete variables, such as Bernoulli or Poisson, specifying the joint distribution of multiple longitudinal measurements incorporating between measurements correlations remains difficult (Molenberghs and Verbeke (2005), Bergsma, Croon and Hagenaars (2009)). As such, jointly modeling the mean, variance, and correlations of repeated discrete measurements is much more challenging, compared with that for continuous cases.

In this paper, we propose a novel approach by using copula for mean-correlation regression analysis for longitudinal zero-inflated data, solving both problems of specifying joint distributions and parsimoniously modeling correlations with no constraint. A copula is a function that represents the joint distribution in terms of its marginals (Sklar (1959)), and hence can be used to couple any discrete and/or continuous distributions. An appealing feature of copula modeling is that it can be used to retain well-known parametric families for the marginal distributions even though they may not be easily extendable to multivariate settings. The use of copulas for count data is not new, for example, Zimmer and Trivedi (2006) use trivariate copula to model two longitudinal count outcomes and a binary outcome. Madsen and Fang (2011) introduced a Gaussian copula likelihood for discrete longitudinal data; Deb, Trivedi and Zimmer (2014) and Shi and Zhang (2015) proposed a copula-based bivariate hurdle model for bivariate outcomes which are a mixture of zeros and continuously measured

positive. Tang, Zhang and Leng (2018) first develop a copula-based mean-correlation modeling approach for classical count data. Zimmer (2018) employed a copula-based method for identifying and estimating the coefficient of a binary endogenous regressor in a Poisson regression. The existing copula-based works mainly use copula to decouple the marginal feature from the dependent structure and generally do not parsimoniously model the correlation or well account for the zero-inflation. We then study the use of hyper-spherical coordinates to parametrize the correlation matrix in the copula in terms of a set of angles, effectively a new set of constraint-free parameters on their support. Aided by this property, we propose separated mean, correlation, and dispersion regression models to understand these three key quantities, which can also handle irregularly and possibly subject-specific times points. We show that our approach is adaptive, flexible, and powerful being innovatively capable of incorporating general covariates in a regression model for correlations, see the Rutgers alcohol data example in Section 3.1.

There is often an issue of computational feasibility arising when maximizing the full likelihood function constructed from the copula representation. The high-dimensional intractable integrals presents substantial challenges in statistical inferences and applications. We propose an inferential strategy based on the pairwise likelihood, which only requires the computation of bivariate distributions, and can guarantee the resulting estimated correlation matrix to be always positive-definite, overcoming an important issue of using the pairwise likelihood approaches for correlation and covariance matrices. The other benefits of our approach are the simplicity of implementation and the potential to handle large data sets. The estimators based on the pairwise likelihood are generally consistent and asymptotically normally distributed. We then demonstrate the usefulness and merits of the proposed framework in terms of simulation and real data example.

The rest of the paper is organized as follows. Section 2 introduces the joint mean-correlation-dispersion modeling approach of the paper and its theoretical properties. Section 3 presents numerical simulation and real data analysis. Conclusions and an outline of future study are found in Section 4. Technical details are relegated to the Appendix.

2 Main methodology

2.1 The joint modeling approach

An appealing approach for incorporating the dependency among longitudinal categorical variables is the copula construction (Joe (1997), Song, Li and Yuan (2009)). The copula approach basically involves the generation of a multivariate joint distribution, given the marginal distributions of the correlated variables, so that the dependence structure is entirely unaffected by the marginal distributions assumed. For our paper, we use the so-called Gaussian copula, which has merits of being convenient and has been demonstrated useful in recent studies (see, e.g., Liu, Lafferty and Wasserman (2009)). Following Sklar (1959), the joint cumulative distribution function (CDF) of random variables $\boldsymbol{U} = (U_1, \ldots, U_d)^{\mathrm{T}}$ with given margins can be constructed by the Gaussian copula in the form

$$G(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d) = \Phi_d(v_1, \ldots, v_d; \mathbf{R}).$$

Here Φ_d is the distribution function of the d-dimensional standardized normal distribution with zero mean, $\mathbf{R}$ is the correlation matrix, and $v_i = \Phi_1^{-1}(w_i)$ where $w_i = P(U_i \leq u_i)$ is the marginal distribution of U_i $(1 \leq i \leq d)$. The copula construction provides substantial flexibility in correlating random variables, as it separates the marginal feature from the dependence structure, and can treat continuous, categorical and mixed data in a unified fashion. Because of the decoupling, models developed for independent data can be seamlessly incorporated by

appropriately manipulating the marginal distributions. Though the Gaussian copula is elaborated in our method, we remark that other copulas can also be applied without comprising the essence of our mean-correlation modeling framework. For example, the t-copula (Fang, Fang and Kotz (2002)) parametrized by the correlation matrix $\mathbf{R}$ may also be applied.

Designate the vector $(Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ to be the m_i longitudinal measurements for the ith subject at times $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})$, with corresponding covariates vectors introduced in the following later. Suppose that Y_{ij} marginally follows the common zero-inflated negative binomial distribution (ZINB):

$$P(Y_{ij} = y) = \begin{cases} p_{ij} + (1 - p_{ij})P(W_{ij} = 0) & y = 0, \\ (1 - p_{ij})P(W_{ij} = y) & y \geq 1, \end{cases} \tag{2.1}$$

where W_{ij} follows a negative binomial distribution with probability mass function

$$P(W_{ij} = w) = \frac{\Gamma(w + 1/\tau)}{w!\Gamma(1/\tau)} \left(\frac{1}{1 + \tau\lambda_{ij}} \right)^{1/\tau} \left(\frac{\tau\lambda_{ij}}{1 + \tau\lambda_{ij}} \right)^{w}, \tag{2.2}$$

where $w = 0, \ldots$ and τ ($\tau > 0$) is a shape parameter that quantifies the amount of overdispersion. Denote $Y_{ij} \sim \mathrm{ZINB}(\lambda_{ij}, p_{ij}; \tau)$, it is easy to verify that the mean and variance of Y_{ij} are $EY_{ij} = (1 - p_{ij})\lambda_{ij}$ and $\mathrm{Var}(Y_{ij}) = (1 - p_{ij})\lambda_{ij}(1 + \tau\lambda_{ij} + p_{ij})$, respectively.

The ZINB model has been well studied for independent zero-inflated data. In a ZINB model, both p_{ij} and λ_{ij} are modeled as functions of explanatory variables. The log link function is used to relate λ_{ij} to the explanatory variables (say, $\mathbf{x}_{ij}$), and the logit link function is used to relate p_{ij} to the explanatory variable (say, h_{ij}). The predictors (say, $\mathbf{x}_{ij}$) for λ_{ij} can be different from the predictors (say, $\mathbf{h}_{ij}$) for p_{ij}. Let us assume that

$$\log(\lambda_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \qquad \mathrm{logit}(p_{ij}) = \mathbf{h}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}. \tag{2.3}$$

Thus, the mean of Y_{ij}, $\mu_{ij} = EY_{ij} = (1 - p_{ij})\lambda_{ij}$, depends on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Let the joint CDF of $Y_{i1}, \ldots, Y_{im_i}$ follow the Gaussian copula representation

$$\mathbf{F}(\mathbf{y}_i) = P(Y_{i1} \leq y_{i1}, \ldots, Y_{im_i} \leq y_{im_i}) = \Phi_{m_i}(z_{i1}, \ldots, z_{im_i}; \mathbf{R}_i), \tag{2.4}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^{\mathrm{T}}$, $z_{ij} = \Phi_1^{-1}\{F_{ij}(y_{ij})\}$ ($j = 1, \ldots, m_i$), $F_{ij}(\cdot)$ is the CDF of Y_{ij} and $\mathbf{R}_i = (\rho_{ijk})_{j,k=1}^{m_i}$ is the correlation matrix of subject i. Clearly, at the model level, the marginal distributions and the correlations of the discrete longitudinal responses are treated separately. Thus this framework provides a powerful and flexible device to incorporate desired marginal models for discrete responses. Note that although the elements in $\mathbf{R}_i$ are not directly the correlations between the discrete observations, they are determining the dependence of the longitudinal observations via (2.4). When the responses are binary, the correlation between two observations is a monotone function of the corresponding element in $\mathbf{R}_i$; see also Fan et al. (2017). We also refer to the discussions in Song (2000) on the connection between the correlation coefficients in $\mathbf{R}_i$ and those of the observed variables.

With so many parameters in $\{\mathbf{R}_i\}$ ($i = 1, \ldots, n$), the model is clearly over-parametrized and thus can not be applied in practice. It is also worth to mention that a regression approach based on a direct Cholesky-type decomposition of the correlation matrix encounters great difficulty. To overcome this, we first decompose $\mathbf{R}_i$ as

$$\mathbf{R}_i = \mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}, \tag{2.5}$$

where $\mathbf{T}_i$ is a lower triangular matrix given by

$$\mathbf{T}_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{i21} & s_{i21} & 0 & \cdots & 0 \\ c_{i31} & c_{i32}s_{i31} & s_{i32}s_{i31} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{im_i1} & c_{im_i2}s_{im_i1} & c_{im_i3}s_{im_i2}s_{im_i1} & \cdots & \prod_{l=1}^{m_i-1} s_{im_il} \end{pmatrix} \tag{2.6}$$

and $c_{ijk} = \cos(\omega_{ijk})$ and $s_{ijk} = \sin(\omega_{ijk})$ are trigonometric functions of angles $\omega_{ijk} \in [0, \pi)$ $(1 \le k < j \le m_i)$ that are the parameters under the new parametrization. That is, the nonzero entries in the lower diagonal matrix $\mathbf{T}_i$ are given by $T_{i11} = 1$, $T_{ij1} = \cos(\omega_{ij1})$ for $j = 2, \ldots, m_i$, and

$$T_{ijk} = \begin{cases} \cos(\omega_{ijk}) \prod_{l=1}^{k-1} \sin(\omega_{ijl}) & 2 \le k < j \le m_i; \\ \prod_{l=1}^{k-1} \sin(\omega_{ijl}) & k = j;\ j = 2, \ldots, m_i. \end{cases} \tag{2.7}$$

Note that for any matrix $\mathbf{T}_i$, $\mathbf{R}_i = \mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}$ is guaranteed to be nonnegative definite. The special form of $\mathbf{T}_i$ in (2.6) ensures further that the diagonals of $\mathbf{R}_i$ are unit and that $\mathbf{R}_i$ is positive definite. In addition, the lower triangular structure of $\mathbf{T}_i$ respects the longitudinal nature of the data in that the angles are added sequentially to $\mathbf{T}_i$ in a way similar to the fact that longitudinal data are collected along a time dimension. Thus, the effect of the decomposition is to transform the unknown positive definite correlations $\{\mathbf{R}_i\}$ into unconstrained parameters in $\{\omega_{ijk}\}$ on $[0, \pi)$. This decomposition in (2.6) appeared in Creal, Koopman and Lucas (2011) for analyzing time series and was studied by Zhang, Leng and Tang (2015) for regression with continuous longitudinal responses where it was argued that the angles ω_{ijk} represent rotations of these coordinates and their magnitude reflects roughly the correlations amongst different components.

Since all angles in (2.6) are unconstrained on $[0, \pi)$, we propose to model these angles $\{\omega_{ijk}\}$ collectively via a regression model after a monotone transformation as

$$\omega_{ijk} = \pi/2 - \operatorname{atan}(\mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{\alpha}), \tag{2.8}$$

where $\mathbf{w}_{ijk} \in \mathbb{R}^q$ is a covariate and $\boldsymbol{\alpha}$ is the $q \times 1$ unknown parameters. We note that ω_{ijk} can be directly modeled as a linear function of $\mathbf{w}_{ijk}$ as in Zhang, Leng and Tang (2015). We remark that $\mathbf{w}_{ijk}$ depends on two indices j and k of the ith subject. This is reasonable since for modeling the correlation between observation j and k, we need to examine the covariates of the ith subject at the two corresponding observations. In practice, we can follow the convention of longitudinal data analysis by taking $\mathbf{w}_{ijk}$ as some function of the time lag $|t_{ij} - t_{ik}|$ between observations, which effectively ensures the correlation to be stationary; see also Pourahmadi (1999). Other time-dependent covariates may also be meaningfully exploited. Such a rationale can be initially assessed by examining empirical correlations from the observed longitudinal data. For a balanced longitudinal study, an initial version of the angles ω_{ijk} can be obtained from the empirical correlation matrix of the $\phi^{-1}(F(y))$ after a marginal model fitting. By examining the plot of those angles ω_{ijk} against the time lag, appropriate models can be used to describe such a curvature. Furthermore, we emphasize that by using regression model (2.8) in conjunction with copula, our approach provides a new device for modeling general joint distributions for data that can be discrete or more generally being mixed.

By combining all unknown parameters in this modeling framework, we write collectively the parameter vector of interest as $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \tau)^{\mathrm{T}}$. Using the ZINB model for the responses marginally in (2.1) and the model in (2.8) for the correlations, we are ready to develop the maximum likelihood estimators for $\boldsymbol{\theta}$. A daunting difficulty is, however, that applying copula to fit discrete data is known to be computationally intensive. To see this, we may write the full likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} P(Y_{i1} = y_{i1}, \ldots, Y_{im_i} = y_{im_i}) \\ &= \prod_{i=1}^{n} P(y_{i1} - 1 < Y_{i1} \le y_{i1}, \ldots, y_{im_i} - 1 < Y_{im_i} \le y_{im_i}) \\ &= \prod_{i=1}^{n} \int \cdots \int_{\mathbf{z}_i^- < \mathbf{u} \le \mathbf{z}_i} \phi_{m_i}(\mathbf{u}; \mathbf{R}_i)\, d\mathbf{u}, \end{aligned} \tag{2.9}$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{im_i})^{\mathrm{T}}$ and $\mathbf{z}_i^- = (z_{i1}^-, \ldots, z_{im_i}^-)^{\mathrm{T}}$ with $z_{ij} = \Phi_1^{-1}\{F_{ij}(y_{ij})\}$, $z_{ij}^- = \Phi_1^{-1}\{F_{ij}(y_{ij} - 1)\}$. When y_{ij} takes the smallest possible value on its support, $z_{ij}^- = -\infty$. The vector inequality $\mathbf{z}_i^- < \mathbf{u} \le \mathbf{z}_i$ means componentwise, that is, $z_{i1}^- < u_1 \le z_{i1}, \ldots, z_{im_1}^- < u_{m_i} \le z_{im_i}$. Though integrals in the full likelihood (2.9) can be approximated numerically or by Bayesian methods, the computational cost is clearly high and may not scale easily to even a moderate number of repeat measurements. Actually, directly calculating the distribution function of each subject i specified by (2.4) requires 2^{m_i} summations of lower dimensional distribution functions as in the approach of Song, Li and Yuan (2009), thus the computational cost grows exponentially with m_i; see also Smith and Khaled (2012).

To overcome the computational difficulty, we propose to apply the composite likelihood idea reviewed in Varin, Reid and Firth (2011) by using pairwise likelihood. The pairwise approach is a good balance between statistical and computational efficiency. Many studies have found that the efficiency loss of the pairwise likelihood estimator (relative to the maximum likelihood estimator) is negligible to small in applications, see, for example, Renard, Molenberghs and Geys (2004), Fieuws and Verbeke (2006).

2.2 Pairwise likelihood (PL) inference

To estimate the parameters in the model specified by (2.3)–(2.8), we apply the composite likelihood idea by constructing the all pairwise likelihood via bivariate copula as

$$pL(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{1 \le j < k \le m_i} \int_{z_{ij}^-}^{z_{ij}} \int_{z_{ik}^-}^{z_{ik}} \phi_2(\mathbf{u}; \rho_{ijk})\, d\mathbf{u}, \tag{2.10}$$

where $\phi_2(\cdot; \rho)$ is the probability density function of bivariate normal $N(0, 0, 1, 1, \rho)$. The computational cost is remarkably lower than that of the full likelihood. To see this, we note that (2.10) involves $m_i(m_i - 1)/2$ summations for each subject in the longitudinal data, a polynomial order complexity as compared to the exponential order in computing the full likelihood. Furthermore, each summand can be obtained by approximating a bivariate normal distribution function which can be evaluated very quickly and accurately with existing computational routines developed for low-dimensional integration, for example, those in Tong (1990) and the ones implemented in R (e.g., function `biv.nt.prob` in package `mnormt`; and function `pmvnorm` in package `mvtnorm`).

By using the pairwise likelihood (2.10) in conjunction with our mean-correlation regression models specified in (2.3)–(2.8), our proposed method also substantially enhances the

conventional pairwise likelihood methods for studying covariance and correlation matrices. We remark that a unique feature of our pairwise likelihood approach is that ρ_{ijk} in (2.10) is specified by the hyperspherical decomposition in (2.6) and (2.8) so that it is highly parsimonious and ensures the resulting correlation matrix to be automatically positive definite. In contrast, a conventional composite pairwise likelihood treats all correlations as standing-alone parameters, ignoring the fact that they are from a correlation matrix. Thus in addition to the difficulty from over-parametrization, the resulting estimates from a conventional pairwise likelihood approach may not respect the fact that the pairwise correlations jointly forms a correlation matrix.

Let the log pair-wise likelihood function be

$$\begin{aligned} \mathrm{pl}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{1 \le j < k \le m_i} \log \int_{z_{ij}^-}^{z_{ij}} \int_{z_{ik}^-}^{z_{ik}} \phi_2(\mathbf{u}; \rho_{ijk})\, d\mathbf{u} \\ &:= \sum_{i=1}^{n} \sum_{1 \le j < k \le m_i} l_{ijk}(\boldsymbol{\theta}), \end{aligned} \tag{2.11}$$

and the score function be

$$S_n(\boldsymbol{\theta}) = \frac{\partial\, \mathrm{pl}}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \sum_{1 \le j < k \le m_i} \frac{\partial l_{ijk}}{\partial \boldsymbol{\theta}} := \sum_{i=1}^{n} S_{ni}(\boldsymbol{\theta}). \tag{2.12}$$

We employ the modified Fisher scoring algorithm to maximize the pairwise likelihood function (2.11). The exact forms of the score function and the expected Hessian matrix for $\mathrm{pl}(\boldsymbol{\theta})$ are provided in the Appendix. Denote $\boldsymbol{\theta}^{(t-1)}$ as the updated value of $\boldsymbol{\theta}$ at the $(t-1)$th iteration. We update the estimates by the following iterative equation $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \mathbf{K}_n^{-1}(\boldsymbol{\theta}^{(t-1)})\mathbf{S}_n(\boldsymbol{\theta}^{(t-1)})$, where $\mathbf{K}_n$ is the expected Hessian matrix given later in (2.13).

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be initialized by fitting the marginal model, we can use the independent correlation structure ($\rho_{ijk} = 0$, thus $\boldsymbol{\alpha} = 0$). These initial estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are known to be root-n consistent (Zeger and Liang (1986)). If data are balanced where $\mathbf{R}_i = \mathbf{R}$, it is not difficult to find an initial consistent estimator of $\boldsymbol{\alpha}$. To do that, we can easily obtain a sample estimator of $\mathbf{R}$ which is root-n consistent, using the initial consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. By noticing $\omega_{1jk} = \cdots = \omega_{njk}$ for balanced data, we can use the model in (2.8) to consistently estimate $\boldsymbol{\alpha}$. It is then straightforward to show that one step estimator will be as efficient as the fully iterated estimators, a reminiscence of what is true for one step estimators for the MLE. If data are unbalanced, obtaining the global optimal solution of the likelihood or the pairwise likelihood is more difficult. We experience, however, that the iterative procedure we have discussed so far always converges to an optimal solution, and the numerical results reported in Section 4 are based on this simple iterative procedure.

2.3 Asymptotic properties

The asymptotic property of the maximum likelihood estimation involves the limit of the expected Hessian matrix $\mathbf{K}(\boldsymbol{\theta}) = \lim_{n\to\infty} -\frac{1}{n} E(\partial^2 \mathrm{pl}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}})$, and the limit of variance $\mathbf{J}(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathrm{Var}_{\boldsymbol{\theta}}(\frac{1}{\sqrt{n}}\mathbf{S}_n(\boldsymbol{\theta}))$, where the expectation is conditioning on the covariates $\mathbf{x}_{ij}$ and $\mathbf{w}_{ijk}$. To formally establish the theoretical properties, we impose the following standard regularity conditions in studying statistical methods for longitudinal data.

Condition A1. The dimensions p, d and q of covariates $\mathbf{x}_{ij}$, $\mathbf{h}_{ij}$ and $\mathbf{w}_{ijk}$ are fixed; $n \to \infty$ and $\max_i m_i$ is bounded.

Condition A2. The true value $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}}, \boldsymbol{\alpha}_0^{\mathrm{T}}, \tau_0)^{\mathrm{T}}$ is in the interior of the parameter space $\boldsymbol{\theta}$ that is a compact subset of $\mathbb{R}^{p+d+q+1}$.

Condition A3. Both $\mathbf{K}(\boldsymbol{\theta}_0)$ and $\mathbf{J}(\boldsymbol{\theta}_0)$ are positive definite matrices.

For the MLE based on the full likelihood function, we have the following asymptotic results.

Theorem 1. *Under regular conditions* (A1) *to* (A3), *let* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}}, \hat{\tau})^{\mathrm{T}}$ *be the maximum pairwise likelihood estimates of* (2.9), *then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\big(0, G(\boldsymbol{\theta}_0)^{-1}\big),$$

where $G(\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{K}(\boldsymbol{\theta})$ *is known as the sandwich information or Godambe information.*

The form of sandwich information is due to the fact, $\mathbf{K}(\boldsymbol{\theta}) \neq \mathbf{J}(\boldsymbol{\theta})$, indicating loss of efficiency with respect to maximum likelihood estimation (Varin, Reid and Firth (2011)). However, Renard, Molenberghs and Geys (2004), Fieuws and Verbeke (2006) among others have found that the efficiency loss of the pairwise likelihood estimator (relative to the maximum likelihood estimator) is negligible to small in applications. Our numerical studies experience in later section is consistent with such conclusion.

Since $\hat{\boldsymbol{\theta}}$ is consistent estimators for $\boldsymbol{\theta}_0$, $\mathbf{K}$ and $\mathbf{J}$ in the asymptotic covariance matrix can be consistently estimated by

$$\mathbf{K}_n(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{1\le j<k\le m_i} \ddot{l}_{ijk}(\hat{\boldsymbol{\theta}}), \tag{2.13}$$

where $\ddot{l}_{ijk}(\boldsymbol{\theta}) = \partial^2 l_{ijk}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}$, and

$$\mathbf{J}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} S_{ni}(\hat{\boldsymbol{\theta}})S_{ni}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}). \tag{2.14}$$

Therefore, $G^{-1}(\boldsymbol{\theta}_0)$ can be consistently estimated by

$$G_n^{-1}(\hat{\boldsymbol{\theta}}) = \mathbf{K}_n^{-1}(\hat{\boldsymbol{\theta}})\mathbf{J}_n(\hat{\boldsymbol{\theta}})\mathbf{K}_n^{-1}(\hat{\boldsymbol{\theta}}). \tag{2.15}$$

3 Examples: Data analysis and simulations

3.1 Rutgers alcohol problem index data

We apply the proposed approach to the dataset on gender differences across two years in alcohol-related problem (Neighbors et al. (2010)), measured by the Rutgers Alcohol Problem Index (RAPI, White and Labouvie (1989)). This dataset is drawn from an intervention study aimed at reducing problematic drinking in college students. After filtering out the objects with incomplete measurements, we have a balanced data set with 2805 longitudinal measures across five time points from 561 individuals, 213 men and 348 women. The histogram of RAPI outcomes show in Figure 1(a) reveals obvious zero inflation. According to Neighbors et al. (2010), we use the zero-inflated negative binominal (ZINB) model with the following links for λ_{it} and p_{it}

$$\log(\lambda_{it}) = \beta_0 + \beta_1 \cdot gender_i + \beta_2 \cdot Time_t + \beta_3 \cdot gender_i \times Time_t, \tag{3.1}$$

$$\operatorname{logit}(p_{it}) = \gamma_0 + \gamma_1 \cdot gender_i + \gamma_2 \cdot Time_t + \gamma_3 \cdot gender_i \times Time_t, \tag{3.2}$$

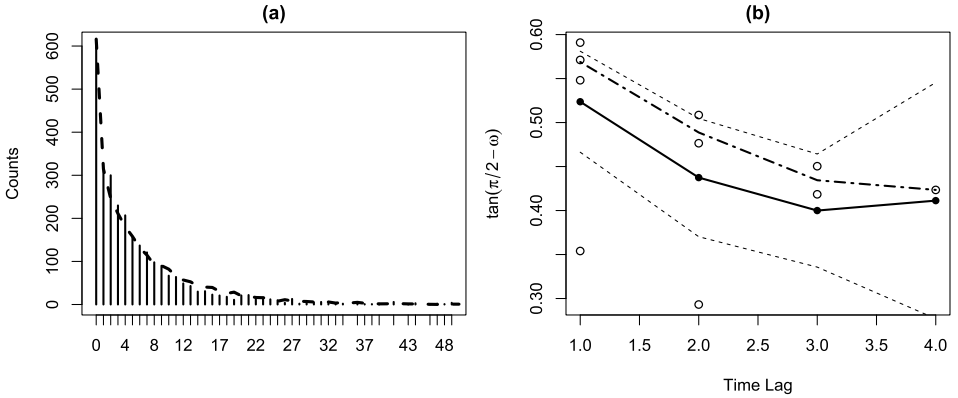

Figure 1 (a) *Histogram of RAPI data and predicted counts* (*dashed lines*); (b) *Plot of fitted correlations versus time lag. Circle dots are fitted angles with a common correlation matrix for all subjects with parametrization* (2.6), *the dashed black line is from fitting a LOWESS curve to the circle dots*, *the solid black line is from the proposed model*, *and the dashed curves represent asymptotic* 95% *confidence intervals*.

As for the correlation modeling, we first investigate a reasonable model using a common 5×5 correlation matrix $\mathbf{R}$ by letting $\mathbf{R}_i = \mathbf{R}$ for all subjects. Thus the equivalent unknown parameters for $\mathbf{R}$ by the decomposition (2.6) are ω_{jk} $(1 \le j < k \le 5)$. Then the pairwise likelihood approach is applied to obtain estimators $\hat{\omega}_{jk}$, leading to an estimated correlation matrix. The plot of the function $\tan(\pi/2 - \hat{\omega}_{jk})$ versus the time lag is given in Figure 1(b) with circle dots, suggesting some monotone decreasing associations. Clearly, this method for incorporating the correlations involves $5 \times 4/2 = 10$ parameters. To model the correlation parsimoniously by regression, we link these angles with covariates via the parsimonious model specified in

$$\tan(\pi/2 - \omega_{ijk}) = \alpha_0 + \alpha_1(t_{ij} - t_{ik}) + \alpha_2(t_{ij} - t_{ik})^2. \tag{3.3}$$

The estimated parameters of the mean-correlation joint model with estimated standard deviation shown in the subscript are $\hat{\beta}_0 = 1.7435_{0.0553}$, $\hat{\beta}_1 = 0.3212_{0.0906}$, $\hat{\beta}_2 = -0.0040_{0.0039}$, $\hat{\beta}_3 = 0.0075_{0.0055}$, suggesting that the *gender* is significant, the covariate *Time* and the interaction effect are marginally significant. For model (3.2), we have $\hat{\gamma}_0 = -4.6255_{0.5686}$, $\hat{\gamma}_1 = -0.1429_{0.9294}$, $\hat{\gamma}_2 = 0.1326_{0.0214}$, $\hat{\gamma}_3 = -0.0039_{0.0358}$, indicating that *Time* is significant. The estimated parameters in the correlation regression model are $\hat{\alpha}_0 = 0.6590_{0.0963}$, $\hat{\alpha}_1 = -0.1600_{0.1052}$, $\hat{\alpha}_2 = 0.0244_{0.0237}$, implying that a reduced model can be further discussed. $\hat{\tau} = 0.1908_{0.0652}$ shows that data are over-dispersed. Denoted by $\hat{\omega}_{jk}$ the estimated angles from the parsimonious model, Figure 1(b) also shows the plot of the fitted angles $\tan(\pi/2 - \hat{\omega}_{jk})$ versus time lag, which indicates a competent fitting of the angles with far fewer parameters where only 3 parameters are involved compared with 10 parameters in a common correlation matrix $\mathbf{R}$. To show the goodness-of-fit of our model, we simulate 1000 new outcomes from the fitted model, then average over simulations to get predicted counts as the black line in Figure 1(a). Figure 1(b) shown the fitted polynomials and the LOWESS curves.

Using ZINB model and pairwise likelihood (2.11), the log pair-wise likelihood value increased from $-31{,}675$ to $-30{,}878$ by modeling the correlation using our approach than assuming independence. After fitting the models, we then simulate 1000 outcomes from the fitted model to compute the average predicted RAPI. Figure 2 shows the average predict RAPI for women and men with comparison to the sample mean of RAPI, respectively, indicating that our proposed model is fitting reasonably well.

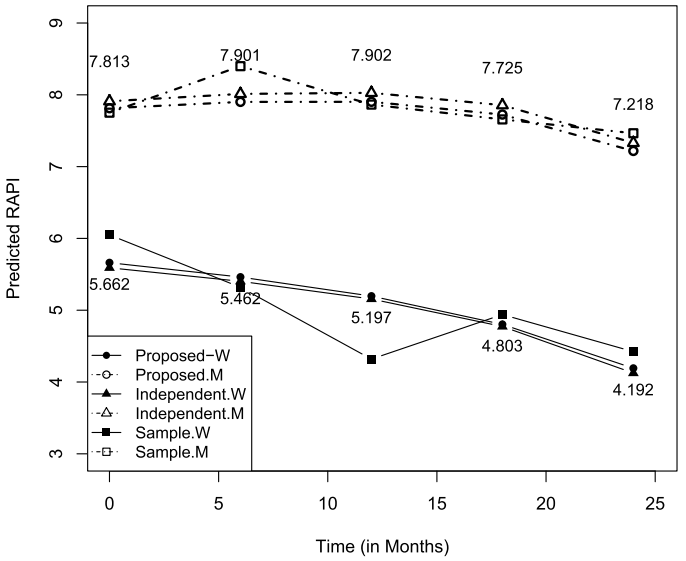

Figure 2 *Plots of Average predicted RAPI for women and men under proposed model, ZINB with assuming independence and sample mean of RAPI.*

3.2 Simulations

In this section, we investigate the finite sample performance of the proposed estimation method and compare our method with the standard ZINB approach. We conduct simulations in two studies. In each of the following studies, we generate 500 data sets and consider sample sizes $n = 200$ and 400. All simulations were conducted in R. The data sets are generated from the model

$$y_{ij} \sim \text{ZINB}(\lambda_{ij}, p_{ij}; \tau), \qquad \log(\lambda_{ij}) = 0.1 + 0.5x_{ij1} - 0.5x_{ij2},$$
$$\text{logit}(p_{ij}) = -1 + 0.5x_{ij1} + 0.5x_{ij2}$$

The angles are then modeled by quadratic polynomials with $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2)^T$ and $\boldsymbol{\alpha} = (1, -1, 0.5)^T$, where the measurement times t_{ij} are generated from the uniform distribution. The shape parameter $\tau = 1/8$. We consider two cases: (I) $m_i \equiv 6$ and (II) $m_i - 1 \sim Binomial(6, 0.8)$, respectively. The latter case gives different numbers of repeated measurements m_i for different subjects. The covariate $x_{ij} = (x_{ij1}, x_{ij2})^T$ is generated from independent standard normal distribution.

To compare our proposed approach with the full likelihood approach and the classical ZINB model (with independence correlation), we directly use the estimation algorithm provided by R package `pscl`.

Tables 1 and 2 show the accuracy of the estimated parameters in terms of their mean biases (MB) and standard deviations. Additionally, to evaluate the inference procedure, we compare the sample standard deviation (SD) of 500 parameter estimates to the sample average of 500 standard errors (SE) using formula (2.15). The standard deviation (Std) of 500 standard errors is also reported. For the parameters β_i in the regression mean, all three approaches give similar mean biases while our approach is more efficient in all the cases. It can also be seen that the SD and SE are quite close especially when n is large, indicating the sandwich estimation formula (2.15) works considerably well. Although estimators based on the pairwise likelihood function is slightly less efficient than the maximum likelihood estimates, they have relatively small biases. In particular, the estimates for the parameters in correlation matrices based on full likelihood approach are highly biased. As discussed earlier, this is likely due to

Table 1 *Simulation results for case I. Mean absolute bias* (*MAB*) *and standard deviation* (*SD*) *of each parameter* (*reported in the subscript*). *SE is the average standard error calculated using the formula* (2.15)

		n					
		Pairwise likelihood		Full likelihood		Independence	
		200	400	200	400	200	400
β_0	MB_{SD}	$-0.001_{0.050}$	$-0.003_{0.032}$	$-0.001_{0.045}$	$-0.003_{0.030}$	$0.001_{0.053}$	$-0.002_{0.035}$
	SE_{std}	$0.031_{0.002}$	$0.022_{0.001}$	–	–	–	–
β_1	MB_{SD}	$0.001_{0.019}$	$0.001_{0.013}$	$0.001_{0.017}$	$-0.001_{0.012}$	$-0.003_{0.027}$	$-0.001_{0.02}$
	SE_{std}	$0.021_{0.002}$	$0.015_{0.001}$	–	–	–	–
β_2	MB_{SD}	$-0.001_{0.021}$	$0.001_{0.014}$	$-0.001_{0.018}$	$0.001_{0.013}$	$0.003_{0.030}$	$0.001_{0.021}$
	SE_{std}	$0.022_{0.002}$	$0.016_{0.001}$	–	–	–	–
γ_0	MB_{SD}	$-0.007_{0.145}$	$-0.002_{0.102}$	$-0.009_{0.136}$	$-0.007_{0.099}$	$-0.009_{0.155}$	$-0.003_{0.107}$
	SE_{std}	$0.118_{0.010}$	$0.084_{0.005}$	–	–	–	–
γ_1	MB_{SD}	$0.004_{0.079}$	$0.005_{0.055}$	$0.004_{0.073}$	$0.007_{0.051}$	$0.002_{0.105}$	$0.005_{0.073}$
	SE_{std}	$0.083_{0.008}$	$0.059_{0.004}$	–	–	–	–
γ_2	MB_{SD}	$0.001_{0.071}$	$0.005_{0.051}$	$0.001_{0.062}$	$0.004_{0.046}$	$-0.005_{0.095}$	$0.007_{0.067}$
	SE_{std}	$0.073_{0.006}$	$0.052_{0.003}$	–	–	–	–
τ	MB_{SD}	$-0.002_{0.025}$	$0.001_{0.020}$	$-0.003_{0.022}$	$-0.001_{0.017}$	$0.003_{0.034}$	$0.002_{0.025}$
	SE_{std}	$0.140_{0.034}$	$0.095_{0.017}$	–	–	–	–
α_0	MB_{SD}	$-0.001_{0.072}$	$-0.003_{0.049}$	$-0.038_{0.061}$	$-0.043_{0.04}$	–	–
	SE_{std}	$0.025_{0.024}$	$0.028_{0.002}$	–	–	–	–
α_1	MB_{SD}	$0.021_{0.410}$	$0.019_{0.277}$	$0.270_{0.303}$	$0.286_{0.205}$	–	–
	SE_{std}	$0.246_{0.113}$	$0.131_{0.107}$	–	–	–	–
α_2	MB_{SD}	$-0.019_{0.468}$	$-0.014_{0.317}$	$-0.300_{0.331}$	$-0.315_{0.223}$	–	–
	SE_{std}	$0.245_{0.132}$	$0.199_{0.117}$	–	–	–	–

the computational difficulty of evaluating multidimensional integrals when a full likelihood is used. Compared to the classical ZINB estimates for estimating the parameters in the mean model, the pairwise likelihood estimates have very competitive performance. Though our method is not designed with specific consideration for enhancing the mean model estimation incorporating correlations from the longitudinal data, we see that their performance is very close to those of the full likelihood and classical ZINB approaches. When the sample size is smaller, the pairwise likelihood estimates even outperform the classical ZINB approach, showing the advantage of using parsimonious correlation models.

4 Conclusion

This paper presents a tool for investigating the correlation of longitudinal zero-inflated count data, which incorporates the dependency through copula construction. By utilizing the unconstrained parametrization of correlation matrix in a copula and a computationally efficient estimation method based on pairwise likelihood, we have shown that our approach for modeling mean-correlation is flexible and allows the development of parametric, nonparametric, semi-parametric models for correlations.

As always in the application of parametric models for data analysis, it is important to determine whether all the necessary model assumptions are valid before performing inference. Checking assumptions on the marginal model specifications can be done similarly as in the classical generalized linear model theory, while assessing the correlation model adequacy is not straightforward especially for unbalanced data. For balanced data, as illustrated in the

Table 2 *Simulation results for case II. Mean absolute bias (MB) and standard deviation (SD) of each parameter (reported in the subscript). SE is the average standard error calculated using the formula (2.15)*

		n					
		Pairwise likelihood		Full likelihood		Independence	
		200	400	200	400	200	400
β_0	MB_{SD}	$-0.004_{0.052}$	$-0.002_{0.036}$	$-0.004_{0.049}$	$-0.002_{0.034}$	$-0.001_{0.054}$	$-0.001_{0.037}$
	SE_{std}	$0.036_{0.003}$	$0.025_{0.002}$	–	–	–	–
β_1	MB_{SD}	$0.001_{0.021}$	$0.001_{0.016}$	$0.001_{0.020}$	$0.001_{0.015}$	$-0.003_{0.031}$	$0.001_{0.021}$
	SE_{std}	$0.025_{0.003}$	$0.018_{0.001}$	–	–	–	–
β_2	MB_{SD}	$0.001_{0.024}$	$0.001_{0.016}$	$-0.001_{0.021}$	$0.001_{0.014}$	$0.003_{0.031}$	$0.001_{0.021}$
	SE_{std}	$0.027_{0.003}$	$0.018_{0.002}$	–	–	–	–
γ_0	MB_{SD}	$-0.001_{0.155}$	$-0.006_{0.108}$	$-0.010_{0.146}$	$-0.01_{0.101}$	$-0.002_{0.158}$	$-0.007_{0.111}$
	SE_{std}	$0.134_{0.013}$	$0.094_{0.006}$	–	–	–	–
γ_1	MB_{SD}	$0.006_{0.090}$	$0.005_{0.066}$	$0.008_{0.085}$	$0.006_{0.064}$	$0.004_{0.115}$	$0.006_{0.080}$
	SE_{std}	$0.097_{0.010}$	$0.068_{0.005}$	–	–	–	–
γ_2	MB_{SD}	$0.004_{0.077}$	$0.001_{0.060}$	$0.004_{0.070}$	$0.001_{0.054}$	$0.001_{0.095}$	$-0.001_{0.075}$
	SE_{std}	$0.085_{0.008}$	$0.059_{0.004}$	–	–	–	–
τ	MB_{SD}	$-0.002_{0.028}$	$-0.002_{0.022}$	$-0.001_{0.026}$	$-0.002_{0.018}$	$0.003_{0.034}$	$0.001_{0.024}$
	SE_{std}	$0.199_{0.051}$	$0.131_{0.027}$	–	–	–	–
α_0	MB_{SD}	$-0.004_{0.090}$	$0.003_{0.065}$	$-0.043_{0.069}$	$-0.04_{0.054}$	–	–
	SE_{std}	$0.028_{0.004}$	$0.02_{0.002}$	–	–	–	–
α_1	MB_{SD}	$0.028_{0.491}$	$-0.023_{0.339}$	$0.293_{0.371}$	$0.266_{0.281}$	–	–
	SE_{std}	$0.260_{0.017}$	$0.041_{0.010}$	–	–	–	–
α_2	MB_{SD}	$-0.031_{0.562}$	$0.029_{0.383}$	$-0.331_{0.429}$	$-0.299_{0.312}$	–	–
	SE_{std}	$0.392_{0.043}$	$0.131_{0.024}$	–	–	–	–

paper, graphical tools to compare the empirical estimates and the model estimates in Figure 1 are useful, counterparts of those are not currently available when data are unbalanced. Other multivariate copulas than the Gaussian copula may also be possible to establish similar frameworks to that presented in this article, but different copula could have significant effect on modeling the covariation. As such, another future line of research is to develop data-driven models for covariations.

Appendix

Computation of score function

Notice that the objective function

$$\text{pl}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{1 \le j < k \le m_i} l_{ijk}(\boldsymbol{\theta}), \tag{A.1}$$

where

$$\begin{aligned} l_{ijk}(\boldsymbol{\theta}) = \log L_{ijk}(\boldsymbol{\theta}) &= \log \int_{z_{ij}^-}^{z_{ij}} \int_{z_{ik}^-}^{z_{ik}} \phi_2(\mathbf{u}; \rho_{ijk})\, d\mathbf{u} \\ &= \log\big(\Phi_2(z_{ij}, z_{ik}; \rho_{ijk}) - \Phi_2\big(z_{ij}^-, z_{ik}; \rho_{ijk}\big) \\ &\quad - \Phi_2\big(z_{ij}, z_{ik}^-; \rho_{ijk}\big) + \Phi_2\big(z_{ij}^-, z_{ik}^-; \rho_{ijk}\big)\big) \end{aligned} \tag{A.2}$$

and $\Phi_2(x, y; \rho)$ is the CDF of bivariate normal $N(0, 0, 1, 1, \rho)$, $z_{ij} = \Phi_1^{-1}\{F(y_{ij})\} = z_{ij}(\boldsymbol{\eta})$, $z_{ij}^- = \Phi_1^{-1}\{F(y_{ij}-1) = z_{ij}^-(\boldsymbol{\eta})\}$, and $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \tau)^{\mathrm{T}}$, we have

$$\begin{aligned}\frac{\partial l_{ijk}}{\partial \boldsymbol{\eta}} &= \frac{1}{L_{ijk}}\frac{\partial L_{ijk}}{\partial \boldsymbol{\eta}} \\ &= \frac{1}{L_{ijk}}\Big(\frac{\partial}{\partial \boldsymbol{\eta}}\Phi_2(z_{ij}, z_{ik}; \rho_{ijk}) - \frac{\partial}{\partial \boldsymbol{\eta}}\Phi_2\big(z_{ij}^-, z_{ik}; \rho_{ijk}\big) \\ &\quad - \frac{\partial}{\partial \boldsymbol{\eta}}\Phi_2\big(z_{ij}, z_{ik}^-; \rho_{ijk}\big) + \frac{\partial}{\partial \boldsymbol{\eta}}\Phi_2\big(z_{ij}^-, z_{ik}^-; \rho_{ijk}\big)\Big),\end{aligned} \tag{A.3}$$

hence we just need to compute the derivative alike

$$\frac{\partial \Phi_2(z_1, z_2; \rho)}{\partial \boldsymbol{\eta}}. \tag{A.4}$$

Infact, we have

$$\begin{aligned}&\frac{\partial \Phi_2(z_1, z_2; \rho)}{\partial \boldsymbol{\eta}} \\ &= \frac{\partial \Phi_2(z_1, z_2; \rho)}{\partial z_1}\frac{\partial z_1}{\partial \boldsymbol{\eta}} + \frac{\partial \Phi_2(z_1, z_2; \rho)}{\partial z_2}\frac{\partial z_2}{\partial \boldsymbol{\eta}} \\ &= \phi(z_1)\Phi_1\Big(\frac{z_2 - \rho z_1}{\sqrt{1-\rho^2}}\Big)\frac{\partial z_1}{\partial \boldsymbol{\eta}} + \phi(z_2)\Phi_1\Big(\frac{z_1 - \rho z_2}{\sqrt{1-\rho^2}}\Big)\frac{\partial z_2}{\partial \boldsymbol{\eta}} \\ &= \Phi_1\Big(\frac{z_2 - \rho z_1}{\sqrt{1-\rho^2}}\Big)\frac{\partial F(y_1)}{\partial \boldsymbol{\eta}} + \Phi_1\Big(\frac{z_1 - \rho z_2}{\sqrt{1-\rho^2}}\Big)\frac{\partial F(y_2)}{\partial \boldsymbol{\eta}},\end{aligned} \tag{A.5}$$

where $z_i = \Phi_1^{-1}\{F(y_i)\}$, $i = 1, 2$. We can finish (A.3) easily by

$$\frac{\partial F(y)}{\partial \boldsymbol{\beta}} = (1-p)\sum_{k=0}^{y} P(W=k)\frac{k-\lambda}{1+\tau\lambda}\mathbf{x}, \tag{A.6}$$

$$\frac{\partial F(y)}{\partial \boldsymbol{\gamma}} = p(1-p)\Big[1 - \sum_{k=0}^{y} P(W=k)\Big]\mathbf{h}, \tag{A.7}$$

$$\frac{\partial F(y)}{\partial \tau} = (1-p)\sum_{k=0}^{y} P(W=k)\frac{t_{k1} + t_{k2} + t_{k3}}{\tau^2}, \tag{A.8}$$

where $t_{k1} = \mathrm{digamma}(\frac{1}{\tau}) - \mathrm{digamma}(k + \frac{1}{\tau})$, $t_{k2} = \log(1+\tau\lambda) - \frac{\tau\lambda}{1+\tau\lambda}$ and $t_{k3} = \frac{k\tau}{1+\tau\lambda}$, with $\mathrm{digamma}(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $P(W=k)$ is the probability mass function of negative binomial distribution in (2.1).

On the other side, Noting that for $j < k$, $\rho_{ijk} = \sum_{s=1}^{j} T_{ijs}T_{iks}$ and

$$\frac{\partial T_{its}}{\partial \boldsymbol{\alpha}} = \begin{cases} T_{its}\Big[-\tan(\omega_{its})\dfrac{\partial \omega_{its}}{\partial \boldsymbol{\alpha}} + \displaystyle\sum_{l=1}^{s-1}\frac{1}{\tan(\omega_{itl})}\frac{\partial \omega_{itl}}{\partial \boldsymbol{\alpha}}\Big] & t > s > 1, \\ T_{its}\displaystyle\sum_{l=1}^{s-1}\frac{1}{\tan(\omega_{itl})}\frac{\partial \omega_{itl}}{\partial \boldsymbol{\alpha}} & t = s > 1, \\ -\sin(\omega_{it1})\dfrac{\partial \omega_{it1}}{\partial \boldsymbol{\alpha}} & s = 1, \end{cases} \tag{A.9}$$

we can obtain the derivative of l_{ijk} respect to $\boldsymbol{\alpha}$ as

$$\begin{aligned}\frac{\partial l_{ijk}}{\partial \boldsymbol{\alpha}} &= \frac{1}{L_{ijk}} \frac{\partial L_{ijk}}{\partial \boldsymbol{\alpha}} \\ &= \frac{1}{L_{ijk}} \big(\phi_2(z_{ij}, z_{ik}; \rho_{ijk}) - \phi_2\big(z_{ij}^{-}, z_{ik}; \rho_{ijk}\big) \\ &\quad - \phi_2\big(z_{ij}, z_{ik}^{-}; \rho_{ijk}\big) + \phi_2\big(z_{ij}^{-}, z_{ik}^{-}; \rho_{ijk}\big)\big) \frac{\partial \rho_{ijk}}{\partial \boldsymbol{\alpha}}.\end{aligned} \tag{A.10}$$

Combining (A.3) and (A.10) leads to the score function $S_n(\boldsymbol{\theta})$.

Proof of the main theorem

The main theorem can be proved by the standard Taylor expansion approach for maximum likelihood estimation as that in Molenberghs and Verbeke (2005) and Tang, Zhang and Leng (2018). Here we give a scratch for easy reference. It is easy to see that $E_{\boldsymbol{\theta}} S_n(\boldsymbol{\theta}) = 0$, thus by Taylor expansion,

$$0 = S_n(\hat{\boldsymbol{\theta}}) = S_n(\boldsymbol{\theta}_0) + \dot{S}_n(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\dot{S}_n = \partial S_n^{\mathrm{T}} / \partial \boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ is within neighborhood of $\boldsymbol{\theta}_0$. Specially, $\tilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ when $n \to \infty$. Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left[-\frac{1}{n} \dot{S}_n(\tilde{\boldsymbol{\theta}}) \right]^{-1} \frac{1}{\sqrt{n}} S_n(\boldsymbol{\theta}_0). \tag{A.11}$$

From Central Limit Theorem, Assumptions A1–A3 and because $E_{\boldsymbol{\theta}_0} S_n(\boldsymbol{\theta}_0) = 0$ and the boundness of $\mathrm{Var}_{\boldsymbol{\theta}_0}(S_{ni}(\boldsymbol{\theta}_0))$, $i = 1, \ldots, n$,

$$\frac{1}{\sqrt{n}} S_n(\boldsymbol{\theta}_0) \to N\big(0, \mathbf{J}(\boldsymbol{\theta}_0)\big). \tag{A.12}$$

By Assumption A3 and Slutsky's theorem, $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal with asymptotic covariance matrix $G(\boldsymbol{\theta}_0)$.

Acknowledgments

We thank the Editor, Associate Editor, and referee for their constructive comments and suggestions that have greatly improved the paper. Zhang and Chen acknowledge support from the National Key Research and Development Plan (No. 2016YFC0800100) and the NSFC of China (No. 11671374, 71771203). All correspondence should be addressed to Chen.

References

Alfo, M. and Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *Canadian Journal of Statistics* **38**, 197–216. MR2682758 https://doi.org/10.1002/cjs.10056

Atkins, D. C. and Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology* **21**, 726–735. https://doi.org/10.1037/0893-3200.21.4.726

Bergsma, W., Croon, M. and Hagenaars, J. A. (2009). *Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data*. Berlin: Springer. MR1057178

Berk, K. N. and Lachenbruch, P. A. (2002). Repeated measures with zeros. *Statistical Methods in Medical Research* **11**, 303–316. https://doi.org/10.1191/0962280202sm293ra

Bulsara, M. K., Holman, C. D. J., Davis, E. A. and Jones, T. W. (2004). Evaluating risk factors associated with severe hypoglycaemia in epidemiology studies—what method should we use? *Diabetic Medicine* **21**, 914–919. https://doi.org/10.1111/j.1464-5491.2004.01250.x

Buu, A., Li, R., Tan, X. and Zucker, R. A. (2012). Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* **31**, 4074–4086. MR3041794 https://doi.org/10.1002/sim.5510

Creal, D., Koopman, S. J. and Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business and Economic Statistics* **29**, 552–563. MR2879242 https://doi.org/10.1198/jbes.2011.10070

Deb, P., Trivedi, P. and Zimmer, D. M. (2014). Cost-offsets of prescription drug expenditures: Data analysis via a copula-based bivariate dynamic hurdle model. *Health Economics* **23**, 1242–1259. https://doi.org/10.1002/hec.2982

Fan, J., Liu, H., Ning, Y. and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society, Series B* **79**, 405–421. MR3611752 https://doi.org/10.1111/rssb.12168

Fang, H., Fang, K. and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis* **82**, 1–16. MR1918612 https://doi.org/10.1006/jmva.2001.2017

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62**, 424–431. MR2227490 https://doi.org/10.1111/j.1541-0420.2006.00507.x

Ghosh, S. K., Mukhopadhyay, P. and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference* **136**, 1360–1375. MR2253768 https://doi.org/10.1016/j.jspi.2004.10.008

Ground, M. and Koch, S. F. (2008). Hurdle models of alcohol and tobacco expenditure in South African households. *The South African Journal of Economics* **76**, 132–143. https://doi.org/10.1111/j.1813-6982.2008.00156.x

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Boca Raton: CRC Press. MR1462613 https://doi.org/10.1201/b13150

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* **30**, 63–77. MR1957361 https://doi.org/10.1080/0266476022000018510

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. New York: Marcel Dekker. MR1169465

Lee, A. H., Wang, K., Scott, J. A., et al (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medicine Research* **15**, 47–61. MR2225145 https://doi.org/10.1191/0962280206sm429oa

Leng, C., Zhang, W. and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association* **105**, 181–193. MR2656048 https://doi.org/10.1198/jasa.2009.tm08485

Lewsey, J. D. and Thomson, W. M. (2004). The utility of the zero-inflated Poisson and zero-inflated negative binomial models: A case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology* **32**, 183–189.

Liu, H., Lafferty, J. D. and Wasserman, L. A. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**, 2295–2328. MR2563983

Liu, M., Zhang, W. and Chen, Y. (2018). Bayesian joint semiparametric mean-covariance modeling for longitudinal data. *Communications in Mathematics and Statistics* **7**, 1–15. MR3611276

Liu, X. and Zhang, W. (2013). A moving average Cholesky factor model in joint mean-covariance modeling for longitudinal data. *Science China. Mathematics* **56**, 2367–2379. MR3123576 https://doi.org/10.1007/s11425-013-4608-y

Madsen, L. and Fang, Y. (2011). Joint regression analysis for discrete longitudinal data. *Biometrics* **67**, 1171–1175. MR2829253 https://doi.org/10.1111/j.1541-0420.2010.01494.x

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling* **5**, 1–19. MR2133525 https://doi.org/10.1191/1471082X05st084oa

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Berlin: Springer. MR2171048

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

Neighbors, C., Lewis, M. A., Atkins, D. C., Jensen, M. M., Walter, T., Fossos, N., et al (2010). Efficacy of web-based personalized normative feedback: A two-year randomized controlled trial. *Journal of Consulting and Clinical Psychology* **78**, 898–911.

Pan, J. and Mackenzie, G. (2003). On modelling mean–covariance structures in longitudinal studies. *Biometrika* **90**, 239–244. MR1966564 https://doi.org/10.1093/biomet/90.1.239

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690. MR1723786 https://doi.org/10.1093/biomet/86.3.677

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–435. MR1782488 https://doi.org/10.1093/biomet/87.2.425

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94**, 1006–1013. MR2376812 https://doi.org/10.1093/biomet/asm073

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* **44**, 649–667. MR2026438 https://doi.org/10.1016/S0167-9473(02)00263-3

Rose, C. E., Martin, S. W., Wannemuehler, K. A. and Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* **16**, 463–481. MR2242134 https://doi.org/10.1080/10543400600719384

Shi, P. and Zhang, W. (2015). Private information in healthcare utilization: Specification of a copula-based hurdle model. *Journal of the Royal Statistical Society Series A Statistics in Society* **178**, 337–361. MR3300007 https://doi.org/10.1111/rssa.12065

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistiques de l'Universite de Paris* **8**, 229–231. MR0125600

Smith, M. S. and Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association* **107**, 290–303. MR2949360 https://doi.org/10.1080/01621459.2011.644501

Song, P. X.-K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65**, 60–68. MR2665846 https://doi.org/10.1111/j.1541-0420.2008.01058.x

Song, P. X. K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27**, 305–320. MR1777506 https://doi.org/10.1111/1467-9469.00191

Tang, C. Y., Zhang, W. and Leng, C. (2018). Discrete longitudinal data modeling with a mean-corrrelation regression approach. *Statistica Sinica*. https://doi.org/10.5705/ss.202016.0435. Preprint. https://doi.org/10.5705/ss.202016.0435

Tong, Y. L. (1990). *The Multivariate Normal Distribution*. Berlin: Springer. MR1029032 https://doi.org/10.1007/978-1-4613-9655-0

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42. MR2796852

White, H. R. and Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol* **50**, 30–37.

Ye, H. and Pan, J. (2006). Modelling covariance structures in generalized estimating equations for longitudinal data. *Biometrika* **93**, 927–941. MR2285080 https://doi.org/10.1093/biomet/93.4.927

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zhang, W. and Leng, C. (2012). A moving average cholesky factor model in covariance modeling for longitudinal data. *Biometrika* **99**, 141–150. MR2899669 https://doi.org/10.1093/biomet/asr068

Zhang, W., Leng, C. and Tang, C. Y. (2015). A joint modeling approach for longitudinal studies. *Journal of the Royal Statistical Society, Series B* **77**, 219–238. MR3299406 https://doi.org/10.1111/rssb.12065

Zimmer, D. (2018). Using copulas to estimate the coefficient of a binary endogenous regressor in a Poisson regression: Application to the effect of insurance on doctor visits. *Health Economics* **27**, 545–556.

Zimmer, D. and Trivedi, P. (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business & Economic Statistics* **24**, 63–76. MR2234712 https://doi.org/10.1198/073500105000000153

Department of Statistics and Finance
School of Management
University of Science and Technology of China
Hefei, Anhui 230026
P.R. China
E-mail: zwp@ustc.edu.cn
wjl91@mail.ustc.edu.cn
xieqf@mail.ustc.edu.cn
cyu@ustc.edu.cn