Conditional generalized estimating equations of mean-variance-correlation for clustered data

Renwen Luo[a], Jianxin Pan[b,*]

[a] *College of Mathematics, Sichuan University, Chengdu 610065, China*
[b] *Department of Mathematics, University of Manchester, Manchester M13 9PL, U.K.*

Abstract

Joint modelling of mean-covariance structure is an important topic in clustered data analysis. Existing methods, such as those based on modified Cholesky decomposition (MCD), alternative Cholesky decomposition (ACD) and hyperspherical coordinates decomposition (HPC), have two main restrictions. First, they often assume that responses in the same cluster are naturally ordered, for example, by time in longitudinal studies. Second, the existing methods model transformed parameters, for instance, the generalized autoregressive parameters and innovation variances in MCD/ACD, and the hyperspherical coordinates in HPC, making the dependence of correlation coefficients or variances on covariates hardly understandable. As an alternative, a data-driven method that models directly the mean, variances and correlation coefficients for clustered data is proposed. Comparing to the existing methods, the proposed approach not only has no need of natural order in responses but also works on original correlation coefficients and variances. The proposed models are flexible and interpretable, and the parameter estimators in joint generalized estimating equations (GEE) are shown to be consistent and asymptotically normally distributed. Consistent model selection criteria in spirit of quasi-likelihood under independence model criterion (QIC) are considered. The use of the proposed approach is demonstrated by intensive simulation studies and real data analysis.

Keywords: Clustered data, Estimator Efficiency, Generalised estimating equation, Mean-variance-correlation models

1. Introduction

Clustered data arise frequently in many fields including public health, geographical sciences, economics and biological sciences. A common type of clustered data is longitudinal data, which consists of repeated measurements on individuals over time. A typical and popular approach to model clustered data is generalized estimating equations (GEE) proposed by Liang and Zeger (1986). GEE has an attractive advantage that the resulting mean parameter estimators are consistent even if the working correlation structure is misspecified (Liang and Zeger, 1986). However, efficiency loss of the estimated mean parameters may arise when variance is misspecified (Wang and Carey, 2003) or within-cluster correlation structure is not correctly identified (Diggle et al., 2002). Furthermore, missing values even make the mean parameter estimators biased when the working covariance matrix structure is misspecified (Daniels and Zhao, 2003). Therefore, it is necessary to have an accurate covariance estimator for reliable statistical inference.

In GEE, variance is assumed to be a known function of mean multiplied by an overdispersion parameter. The overdispersion and correlation are then estimated by residual moments methods (Liang and Zeger, 1986). However, even in some simple cases of correlation misspecification, the residual moments estimators may be not available (Crowder, 1995). Thus certain joint modelling approaches for mean and covariance for clustered data were studied. For example, a modified Cholesky decomposition (MCD) based method was developed to model the mean-covariance structure for longitudinal data, see, e.g., Pourahmadi (1999); Pan and Mackenzie (2003); Ye and Pan (2006); Leng et al. (2010). Beside MCD, hyperspherical coordinates decomposition (HPC) was proposed by Zhang et al. (2015) with geometric intuition for longitudinal data. The aforementioned approaches also provide good within-subject correlation interpretation in terms of time series (Fan et al., 2007). In fact, MCD and HPC are very useful in joint modelling of mean and covariance for longitudinal data. However, these approaches intrinsically assume that responses in a cluster are naturally ordered, which may be violated in practice, for example, in spatial data analysis. In addition, in real data analysis statistical meanings of the estimated parameters based on MCD and HPC are often not easily

*Corresponding author
Email address: jianxin.pan@manchester.ac.uk (Jianxin Pan)

interpretable in terms of the original correlation coefficients and variances. Also, the angle coordinates in HPC approach are very difficult to interpret to practitioners in many scientific fields.

To overcome the drawbacks of MCD and HPC, we propose a joint mean-variance-correlation (JMVC) model which does not use any matrix decomposition nor involve transformation/reparameterization of variances and correlation coefficients. In fact, it models the mean, variances and correlation coefficients directly. Thus meanings of the estimated parameters in JMVC are interpretable, and also the proposed approach does not require responses within cluster to be ordered, implying that the proposed method is applicable to any correlated data as long as the mean, variances and correlation coefficients are of concern. The resulting estimators of three sets of parameters are roots of three estimating equations, of which the first one is exactly generalized estimating equation for the mean parameters and the other two are conditional generalized estimating equations for the parameters in variances and correlation coefficients, where the conditional generalized estimating equations mean that when estimating one set of parameters, the other sets of parameters are fixed. The parameter estimators are shown to be consistent and asymptotically normally distributed. Compared to the standard GEE approach, the proposed method improves the estimation efficiency of the mean parameters because the covariance structure is correctly modeled. In addition, a new model selection criterion in spirit of quasi-likelihood under independence model criterion (QIC) is proposed and its implementation with efficient computational strategy for selecting the best model is considered. The consistency of the computational strategy is rigorously established and can be regarded as a great improvement of the strategy in Zhang (2012), where they considered the selection consistency of a similar algorithm under normality assumption. In contrast, our approach relaxes the normality assumption and only requires the existence of the first four moments of responses.

This paper is organized as follows. In section 2, we focus on model, estimation procedure and computational algorithm. Theoretical properties of the proposed parameter estimators are studied in section 3. Model selection strategy with an efficient search algorithm is provided in section 4. Section 5 presents intensive numerical simulation studies, which confirms the advantage of the proposed approach. Real data analysis is conducted

in section 6. A concluding summary of main findings and future interests is presented in section 7. Technical proofs of theoretical properties are provided in Appendix and the supplementary materials.

2. Methodology

2.1. Notations

We first introduce some notations. Let y_{ij} be the jth observation of m_i measurements on the ith of n clusters. Denote by $y_i = (y_{i1}, y_{i2}, ...y_{im_i})^T$ the $m_i \times 1$ vector of responses. Suppose $\mathrm{E}(y_i) = \mu_i = (\mu_{i1}, ..., \mu_{im_i})^T$ is $m_i \times 1$ vector of mean of y_i. Denote by $X_i = (x_{i1}, ..., x_{im_i})^T$ the $m_i \times p$ the design matrix with $p \times 1$ covariate x_{ij}. By allowing m_i to be cluster specific, our approach is valid for unbalanced clustered data. Let $\sigma_{ij}^2 = \mathrm{var}(y_{ij})$ be the variance of y_{ij} and $\rho_{ijk} = \mathrm{corr}(y_{ij}, y_{ik})$ be the correlation between y_{ij} and y_{ik}. Let $\mathrm{var}(y_i) = \Sigma_i$ be $m_i \times m_i$ covariance matrix of y_i. It follows that $\Sigma_i = D_i R_i D_i$, where $D_i = diag\{\sigma_{ij}^2, ..., \sigma_{im_i}^2\}$ and $R_i = (\rho_{ijk})_{j,k=1}^{m_i}$ is the correlation matrix for y_i. Let $\epsilon_{ij} = y_{ij} - \mu_{ij}$ and $\delta_{ijk} = \epsilon_{ij}\epsilon_{ik}/\sigma_{ij}\sigma_{ik}$, it follows that $E(\epsilon_{ij}^2) = \sigma_{ij}^2$ and $E(\delta_{ijk}) = \rho_{ijk}$. Let $\epsilon_i^2 = (\epsilon_{i1}^2, ..., \epsilon_{im_i}^2)^T$, $\delta_i = (\delta_{i12}, \delta_{i13}, ..., \delta_{i1m_i}, \delta_{i23}, ..., \delta_{i2m_i}, ..., \delta_{im_{i-1}m_i})^T$. Suppose $\sigma_i^2 = E(\epsilon_i^2)$ and $\rho_i = E(\delta_i)$. For a $n \times n$ matrix $A = (a_{ij})_{i,j=1}^n$, denote $\mathrm{vech}(A)$ by the vector which vectorizes the lower triangular elements A through column by column but does not include the main diagonal elements. It follows that $\rho_i = \mathrm{vech}(R_i)$. Let $\|A\| = \lambda_{\max}(A^T A)^{1/2}$ be matrix spectral norm, where $\lambda_{\max}$ denotes the maximum eigenvalue. Denote $\mathrm{tr}(A)$ by trace of A. For any vector $a = (a_1, ..., a_n)^T$, let $\|a\|$ be Euclid norm. Finally, denote $O_p(M)$ and $o_p(M)$ by the quantities such that $O_p(M)/M$ is bounded in probability and $o_p(M)/M \to 0$ in probability.

2.2. Brief review of generalized estimating equation

For clear exposition of our approach, the conventional GEE approach (Liang and Zeger, 1986) is briefly reviewed first. The GEE approach assumes that marginal mean μ_{ij} and associated covariates x_{ij} are linked to each other through a link function $g(\cdot)$ such that $g(\mu_{ij}) = x_{ij}^T \beta$. And the marginal variance σ_{ij}^2 is a function of mean μ_{ij}, that is, $\sigma_{ij}^2 = \phi_{ij} v(\mu_{ij})$, where ϕ_{ij} is a dispersion parameter and $v(\cdot)$ is a known function. Then

the root $\hat{\beta}_G$ of the equation

$$G(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T D_i^{-1/2} R_i^{-1} D_i^{-1/2} \Big(y_i - \mu_i(X_i \beta) \Big) = 0 \tag{1}$$

is the GEE estimator of parameter β, where $\partial \mu_i / \partial \beta$ is $m_i \times p$ matrix with jth row $\partial \mu_{ij} / \partial \beta = x_{ij}^T \dot{g}^{-1}(x_{ij}^T \beta)$ in which $\dot{g}^{-1}(\cdot)$ is the derivative of $g^{-1}(\cdot)$. In (1), the dispersion parameter ϕ_{ij} and correlation matrix R_i are estimated by residual moments methods (Liang and Zeger, 1986). Note that for Gaussian data, the variance is exactly the overdispersion parameter, that is, $\sigma_{ij}^2 = \phi_{ij}$.

In (1), the GEE estimator $\hat{\beta}_G$ is consistent even if the working correlation is misspecified. Such a misspecification, however, may lead to efficiency loss (Wang and Carey, 2003; Diggle et al., 2002). Liang and Zeger (1986) proposed to use working correlation matrix to avoid such misspecification. However, Crowder (1995) pointed that the residual moments estimators may not exist even in some simple cases. For example, suppose the true correlation structure is equicorrelated, $(R_i)_{jk} = \rho$, and that the working correlation structure is Order-1 Autoregressive, $(R_i)_{jk} = \alpha^{|j-k|}$. Crowder (1995) proved that in such case the there may be no general asymptotic theory supporting the existence and consistency of the residual moments estimator $\hat{\alpha}$.

2.3. Joint mean-variance-correlation model

Now we present our model. The mean, variance and correlation coefficient for clustered data are jointly modeled by

$$g(\mu_{ij}) = x_{ij}^T \beta \quad log(\sigma_{ij}^2) = z_{ij}^T \lambda \quad f(\rho_{ijk}) = h_{ijk}^T \gamma \tag{2}$$

where the dimensions of associated covariates x_{ij}, z_{ij} and h_{ijk} are p, q and d, respectively. λ and γ are parameters corresponding to σ_{ij}^2 and ρ_{ijk}, $g(\cdot)$ is a monotonic and differentiable function linking μ_{ij} and $x_{ij}^T \beta$, and $f(t) = log((1+t)/(1-t))$ is the Fisher-transformation mapping ρ_{ijk} from $(-1, 1)$ to $(-\infty, +\infty)$, which ensures the estimated $\hat{\rho}_{ijk}$ are well defined.

The three equations in (2) are known as joint mean-variance-correlation (JMVC) models. The idea of JMVC is to treat the variance and the correlation as equally important as the mean when modelling clustered data. As mentioned in introduction, MCD,

ACD and HPC often assume that responses in the same cluster are naturally ordered. In addition, they model transformed parameters, making the dependence of correlation coefficients or variances on covariates hardly understandable. In contrast to these existing methods, we directly model the variance and correlation so that the meaning of the estimated parameters are interpretable and our model does not require natural order in cluster while both MCD and HPC do. After fitting JMVC, the correct covariance structure is identified and therefore we expect our model can improve the estimation efficiency over the convention GEE, which is confirmed by our simulation studies.

2.4. Conditional estimating equations of JMVC model

In this section, we present three estimating equations to estimate β, λ, γ in (2). Recall that $\sigma_i^2 = (\sigma_{i1}^2, ..., \sigma_{im_i}^2)^T$ and $\rho_i = vech(R_i)$. We propose the following three estimating equations:

$$\begin{aligned} S_1(\beta) &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T \Sigma_i^{-1}(y_i - \mu_i) = 0 \\ S_2(\lambda)\big|_{\beta=\hat{\beta}} &= \sum_{i=1}^{n} \left(\frac{\partial \sigma_i^2}{\partial \lambda}\right)^T W_i^{-1}(\hat{\epsilon}_i^2 - \sigma_i^2) = 0 \\ S_3(\gamma)\big|_{\beta=\hat{\beta},\lambda=\hat{\lambda}} &= \sum_{i=1}^{n} \left(\frac{\partial \rho_i}{\partial \gamma}\right)^T V_i^{-1}(\hat{\delta}_i - \rho_i) = 0 \end{aligned} \tag{3}$$

where $\hat{\beta}$ and $\hat{\lambda}$ are the solutions of the first and second equations in (3), respectively. In addition,

$$\begin{aligned} \hat{\epsilon}_i^2 &= (\epsilon_{i1}^2(\hat{\beta}), ..., \epsilon_{im_i}^2(\hat{\beta}))^T \\ \hat{\delta}_i &= (\delta_{i12}(\hat{\xi}), ..., \delta_{i1m_i}(\hat{\xi}), \delta_{i23}(\hat{\xi}), ..., \delta_{i2m_i}(\hat{\xi}), ..., \delta_{im_{i-1}m_i}(\hat{\xi}))^T \end{aligned}$$

where $\hat{\xi} = (\hat{\beta}^T, \hat{\lambda}^T)^T$. $Z_i = (z_{i1}, ..., z_{im_i})^T$ and $H_i = (h_{i12}, ..., h_{i1m_i}, ..., h_{im_{i-1}m_i})^T$ are the associated design matrices. Σ_i, W_i and V_i are the covariance matrices of y_i, ϵ_i^2 and δ_i, respectively. $\partial\sigma_i^2/\partial\lambda$ is $m_i \times d$ matrix with jth row $\sigma_{ij}^2 z_{ij}^T$, $\partial\delta_i/\partial\gamma$ is $m_i(m_i-1)/2 \times d$ matrix with $(j-1)*(2m_i-j)/2+k-j$th row $\dot{f}^{-1}(h_{ijk}^T\gamma)h_{ijk}^T$. Note that the second estimating equation involves $\hat{\beta}$ in $\hat{\epsilon}_i^2$ and the third estimating equation involves $\hat{\xi}$ in $\hat{\delta}_i$.

Compared to (3), the generalized estimating equations below

$$S_2(\lambda) = \sum_{i=1}^{n} \left(\frac{\partial \sigma_i^2}{\partial \lambda} \right)^T W_i^{-1} (\epsilon_i^2 - \sigma_i^2) = 0$$
$$S_3(\gamma) = \sum_{i=1}^{n} \left(\frac{\partial \rho_i}{\partial \gamma} \right)^T V_i^{-1} (\delta_i - \rho_i) = 0$$

are the standard GEEs. The second and third conditional estimating equations in (3) are inspired by the consistency of the GEE estimator. Specifically, even though the covariance Σ_i is misspecified, we can obtain consistent $\hat{\beta}$. Replacing β in $S_2(\lambda)$ by this consistent $\hat{\beta}$, we expect to obtain consistent estimator $\hat{\lambda}$. Also, by replacing (β, λ) in $S_3(\gamma)$ by consistent $(\hat{\beta}, \hat{\lambda})$, we expect to obtain consistent estimator $\hat{\gamma}$. Therefore it is reasonable to use the conditional estimating equations $S_2(\lambda)\big|_{\beta=\hat{\beta}}$ and $S_3(\gamma)\big|_{\xi=\hat{\xi}}$ to obtain estimators of λ and γ. Note it is likely that $\mathrm{E}[S_2(\lambda)\big|_{\beta=\hat{\beta}}] \neq 0$ and $\mathrm{E}[S_3(\gamma)\big|_{\xi=\hat{\xi}}] \neq 0$ so that the conditional GEEs $S_2(\lambda)\big|_{\beta=\hat{\beta}} = 0$ and $S_3(\gamma)\big|_{\beta=\hat{\beta}} = 0$ are not exactly the generalized estimating equations for λ and γ. However, we show in supplementary material that as $n \to \infty$, they are asymptotically the same as the generalized estimating equations.

Note that W_i and V_i should be specified. When y_i follows multivariate normal distribution $N_{m_i}(\mu_i, \Sigma_i)$, by some calculations presented in Appendix A we find that $\mathrm{cov}(\epsilon_{ij}^2, \epsilon_{ik}^2) = 2\rho_{ijk}^2 \sigma_{ij}^2 \sigma_{ik}^2$ and $\mathrm{cov}(\delta_{ijk}, \delta_{ilm}) = \rho_{ijl}\rho_{ikm} + \rho_{ijm}\rho_{ikl}$, which indicates that $var(\epsilon_{ij}^2) = 2\sigma_{ij}^4$ and $var(\delta_{ijk}) = 1 + \rho_{ijk}^2$. When the assumption that y_i follows normal distribution is violated, the expressions for elements of W_i and V_i, however, are analytically intractable. For such a reason, in spirit of the idea of Ye and Pan (2006), we approximate W_i and V_i by using the following matrices

$$\widetilde{W_i} = P_{i1}^{\frac{1}{2}} R_{i1}(u_1) P_{i1}^{\frac{1}{2}}, \quad \widetilde{V_i} = P_{i2}^{\frac{1}{2}} R_{i2}(u_2) P_{i2}^{\frac{1}{2}}$$

where

$$P_{i1} = diag(2\sigma_{i1}^4, ..., 2\sigma_{im_i}^4), \quad P_{i2} = diag(1 + \rho_{i12}^2, ..., 1 + \rho_{i1m_i}^2, ..., 1 + \rho_{m_i-1m_i}^2) \tag{4}$$

$R_{i1}(u_1)$ and $R_{i1}(u_1)$ are working correlation matrices, which often take the Compound Symmetry (CS) structure or the Order-1 Autoregressive (AR(1)) structure for longitudinal data. They are of course approximations to the true correlation matrices of ϵ_i^2 and δ_i, respectively.

Algorithm 1: Conditional GEEs for clustered data

1: Input an initial $\beta^{(0)}$, $\lambda^{(0)}$ and $\gamma^{(0)}$, set $k = 0$.

2: Given $\beta^{(k)}$, in particular given $\hat{\epsilon}_i^2$, choose $\lambda^{(k)}$ as initial values, update λ by

$$\lambda^{(s+1)} = \lambda^{(s)} + \left\{ \left[\sum_{i=1}^{n} \left(\frac{\partial \sigma_i^2}{\partial \lambda} \right)^T \widetilde{W}_i^{-1} \left(\frac{\partial \sigma_i^2}{\partial \lambda} \right) \right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial \sigma_i^2}{\partial \lambda} \right) \widetilde{W}_i^{-1} (\hat{\epsilon}_i^2 - \sigma_i^2) \right] \right\} \Bigg|_{\beta=\beta^{(k)}, \lambda=\lambda^{(s)}}$$

until convergence. Denote the result by $\lambda^{(k+1)}$.

3: Given $\beta^{(k)}$ and $\lambda^{(k+1)}$, in particular given $\hat{\delta}_i$, choose $\gamma^{(k)}$ as initial values, update γ by

$$\gamma^{(s+1)} = \gamma^{(s)} + \left\{ \left[\sum_{i=1}^{n} \left(\frac{\partial \delta_i}{\partial \gamma} \right)^T \widetilde{V}_i^{-1} \left(\frac{\partial \delta_i}{\partial \gamma} \right) \right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial \delta_i}{\partial \gamma} \right) \widetilde{V}_i^{-1} (\hat{\delta}_i - \rho_i) \right] \right\} \Bigg|_{\beta=\beta^{(k)}, \lambda=\lambda^{(k+1)}, \gamma=\gamma^{(s)}}$$

until convergence and denote the result by $\gamma^{(k+1)}$.

4: Given $\lambda^{(k+1)}$ and $\gamma^{(k+1)}$, in particular given Σ_i, update β by

$$\beta^{(k+1)} = \beta^{(k)} + \left\{ \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T \Sigma_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right) \Sigma_i^{-1} (y_i - \mu_i) \right] \right\} \Bigg|_{\beta=\beta^{(k)}, \lambda=\lambda^{(k+1)}, \gamma=\gamma^{(k+1)}}$$

5: Replace $\beta^{(k)}$, $\lambda^{(k)}$ and $\gamma^{(k)}$ by $\beta^{(k+1)}$, $\lambda^{(k+1)}$ and $\gamma^{(k+1)}$, respectively. Repeat steps 2-4 until a convergence criterion is met.

We define the JMVC estimators $(\hat{\beta}^T, \hat{\lambda}^T, \hat{\gamma}^T)^T$ as the roots of three estimating equations in (3). As mentioned in section 1, the advantage of GEE is the resulting estimators are consistent even if the correlation matrix is misspecified. Thus we anticipate that not only GEE estimator $\hat{\beta}$, but also conditional GEE estimators $\hat{\lambda}$ and $\hat{\gamma}$ are consistent, which is presented in Theorem 1 and confirmed by our simulation studies in later section. Indeed, any change of nuisance parameters u_1 and u_2 or working correlation structures of R_{1i} and R_{2i} has little effect on the estimators of λ and γ.

We next provide iterative quasi-Fisher algorithm to solve the three estimating equations in (3), which is summarized in Algorithm 1.

Note that the initial values $\beta^{(0)}$, $\lambda^{(0)}$ and $\gamma^{(0)}$ should be given properly. It is natural to use any $\sqrt{n}$-consistent estimate of β as initial value of β. Thus the conventional generalized estimating equation estimator $\hat{\beta}_G$ can be set as $\beta^{(0)}$. To obtain $\lambda^{(0)}$ and $\gamma^{(0)}$, one may first use Algorithm 1 to calculate reasonable values of λ and γ by setting the initial values of λ and γ as the vectors of zeros, then take these two reasonable values as $\lambda^{(0)}$ and $\gamma^{(0)}$.

3. Asymptotic Property

In this section, we present the consistency and asymptotic normality of the JMVC estimators $(\hat{\beta}_n^T, \hat{\lambda}_n^T, \hat{\gamma}_n^T)^T$ under certain regularity conditions presented in supplementary materials. In addition, we also prove that for any subject i, the probability of the estimated correlation matrix $\hat{R}_i$ to be positive definite tends to 1 as $n \to \infty$, under the condition that the true correlation matrix R_i of y_i is positive definite. The main results are summarized as follows.

Theorem 1. *Under regularity conditions C1-C4 presented in supplementary materials, the JMVC estimators* $\hat{\theta}_n = (\hat{\beta}_n^T, \hat{\lambda}_n^T, \hat{\gamma}_n^T)^T$ *are* $\sqrt{n}$*-consistent, that is,* $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$

Theorem 2. *Under regularity conditions C1-C7 presented in supplementary materials, the JMVC estimators* $\hat{\beta}_n$, $\hat{\lambda}_n$ *and* $\hat{\gamma}_n$ *are asymptotically distributed with*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, v_{11}^{-1}),$$

conditional on $\hat{\beta}_n$,

$$\sqrt{n}(\hat{\lambda}_n - \lambda_0)|_{\beta=\hat{\beta}_n} \xrightarrow{\mathcal{D}} N(\phi_n^{\lambda}, v_{22}^{-1}),$$

and conditional on $\hat{\xi}_n = (\hat{\beta}_n^T, \hat{\lambda}_n^T)^T$

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0)|_{\xi=\hat{\xi}_n} \xrightarrow{\mathcal{D}} N(\phi_n^{\gamma}, v_{33}^{-1}),$$

where

$$v_{11} = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T \Sigma_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right)\right],$$

$$v_{22} = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T W_i^{-1}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)\right],$$

$$v_{33} = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T V_i^{-1}\left(\frac{\partial\rho_i}{\partial\lambda}\right)\right],$$

and

$$\phi_n^{\lambda} = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T W_i^{-1}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T W_i^{-1}\left(\frac{\partial\epsilon_i^2}{\partial\beta}\right)(\hat{\beta}_n - \beta_0)\right] = O_p(1),$$

$$\phi_n^{\gamma} = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T V_i^{-1}\left(\frac{\partial\rho_i}{\partial\gamma}\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T V_i^{-1}\left(\frac{\partial\delta_i}{\partial\xi}\right)(\hat{\xi}_n - \xi_0)\right] = O_p(1),$$

Theorem 3. *Under regularity conditions C1-C4 presented in supplementary materials, if the true correlation matrix* R_i *is positive definite, we have*

$$Pr(\eta_i^T \hat{R}_i \eta_i > 0) \longrightarrow 1$$

as $n \to \infty$, *where* $\eta_i \in R^{m_i}$, $i = 1, ..., n$.

The proofs of the above three theorems are provided in supplementary materials. In the proof of Theorem 1, we show that each of two conditional estimating equations can be divided into two terms, in which the first term is a generalized estimating equation and the second term is negligible as long as n is sufficient large, indicating that each of two conditional estimating equations is asymptotically equivalent to a generalized estimating equation. Note that the first asymptotic distribution in Theorem 2 is marginal distribution because the estimating equation for β is standard GEE. And Theorem 2 indicates that $\sqrt{n}\hat{\lambda}_n$ and $\sqrt{n}\hat{\gamma}_n$ have asymptotic biases of scale $O_p(1)$. The reason is that we use conditional generalized estimating equations. The biases of $\hat{\lambda}_n$ and $\hat{\gamma}_n$, i.e., $\phi_n^{\lambda}/\sqrt{n}$ and $\phi_n^{\gamma}/\sqrt{n}$, then converge to zero as $n \to \infty$. Theorem 3 is actually a direct extension of Theorem 1. By consistency of estimator $\hat{\gamma}_n$, the estimated correlation matrix converges to the true correlation matrix, thus also is positive definite as $n \to \infty$.

The covariance of $\hat{\beta}_n$ can be estimated using sandwich formula.

$$\mathrm{cov}(\hat{\beta}_n) = (nv_n^{11})^{-1}(nv_n^{11})(nv_n^{11})^{-1}|_{\theta=\hat{\theta}} = \left[\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T \Sigma_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}} \tag{5}$$

where v_n^{11} is covariance of $S_1(\beta)/\sqrt{n}$. The covariance of $\hat{\lambda}_n$ and $\hat{\gamma}_n$ can be estimated in a similar way. We discuss the explicit expressions in Appendix B. The simulation studies show that the proposed sandwich formulas perform very well.

4. Model Selection

In this section, we present, the model selection method combining the quasi-likelihood under the independence model criterion (QIC) (Pan, 2001) and Bayesian information criterion (BIC) (Schwarz, 1978) within quasi-likelihood framework.

Pan (2001) proposed QIC as:

$$QIC(\hat{\beta}_G; I) = \frac{1}{n}\Big[-2\mathrm{Q}(\hat{\beta}_G; I) + 2\mathrm{tr}(\hat{\Omega}_I\hat{V}_r)\Big]$$

for the model selection of generalized estimating equation (1), where the identity covariance matrix I indicates the quasi-likelihood

$$Q(\beta; I) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} Q(\beta; Y_{ij}, x_{ij})$$

with

$$Q(\beta; y, x) = \int_y^{\mu} \frac{y-t}{\phi V(t)}dt,$$

$\hat{\Omega}_I = -\partial^2 Q(\hat{\beta}_G; I)/\partial\beta\beta^T$ and $\hat{V}_r$ is an estimator of $cov(\hat{\beta}_G)$ and can be obtained by sandwich covariance formula in Liang and Zeger (1986). Here $\mathrm{tr}(\hat{\Omega}_I\hat{V}_r)$ plays the same role as the degree of freedom of parameter. Pan (2001) pointed out that QIC is based on Akaike information criterion (AIC) (Akaike, 1998), in particular, the quasi-likelihood in QIC plays the same role as the log-likelihood in AIC. However, it is well known that AIC has a tendency to select an overparameterized model. Another selection criterion, BIC, penalizes free parameters more strongly and often selects a more parsimonious model. Therefore a natural idea is to modify QIC based on BIC as

$$QIC_m(\hat{\beta}_G; I) = \frac{1}{n}\Big[-2Q(\hat{\beta}_G; I) + log(n)\mathrm{tr}(\hat{\Omega}_I\hat{V}_r)\Big].$$

Inspired by this modification, we now propose new model selection criteria for JMVC. Denote $\mathcal{M} = (\mathcal{M}_\beta, \mathcal{M}_\lambda, \mathcal{M}_\gamma)$ by an arbitrary candidate model where $\mathcal{M}_\beta = \{j_1, ..., j_{p^*}\}$ includes $\{X_{ij_1}, ..., X_{ij_{p^*}}\}$ as the relevant predictors in the model of the mean, where X_{ij} is jth colum in X_i; $\mathcal{M}_\lambda = \{k_1, ..., k_{q^*}\}$ and $\mathcal{M}_\gamma = \{l_1, ..., l_{d^*}\}$ are defined similarly. Denote the true model by $\mathcal{M}^o = (\mathcal{M}^o_\beta, \mathcal{M}^o_\lambda, \mathcal{M}^o_\gamma)$. We define the family of overfitted models as $\mathrm{M}^+ = (\mathrm{M}^+_\beta, \mathrm{M}^+_\lambda, \mathrm{M}^+_\gamma)$ and that of underfitted models as $\mathrm{M}^- = (\mathrm{M}^-_\beta, \mathrm{M}^-_\lambda, \mathrm{M}^-_\gamma)$. Thus if for any $\mathcal{M}_\beta \in \mathrm{M}^+_\beta$, we then have $\mathcal{M}^o_\beta \subset \mathcal{M}_\beta$ and if for any $\mathcal{M}_\beta \in \mathrm{M}^-_\beta$, we must have $\mathcal{M}^o_\beta \not\subset \mathcal{M}_\beta$. Denote the saturated model by $\mathcal{M}^s = (\mathcal{M}^s_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma)$, where $\mathcal{M}^s_\beta = \{1, ..., p\}$, $\mathcal{M}^s_\lambda = \{1, ..., q\}$ and $\mathcal{M}^s_\gamma = \{1, ..., d\}$. Denote $\beta_{\mathcal{M}_\beta} = (\beta_{j_1}, ..., \beta_{j_{p^*}})^T$ and similarly $\lambda_{\mathcal{M}_\lambda}$ and $\gamma_{\mathcal{M}_\gamma}$. With these notations, the model selection criteria for JMVC are proposed as follows

$$
\begin{aligned}
QIC^\beta_{JMVC}(\mathcal{M}_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma) &= \frac{1}{n}\left[-2Q_\beta(\hat\beta_{\mathcal{M}_\beta}, \hat\lambda_{\mathcal{M}^s_\lambda}, \hat\gamma_{\mathcal{M}^s_\gamma}; I) + log(n)\mathrm{tr}(\hat\Omega_\beta \hat V_\beta)\right], \\
QIC^\lambda_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}_\lambda, \mathcal{M}^s_\gamma) &= \frac{1}{n}\left[-2Q_\lambda(\hat\beta_{\mathcal{M}^s_\beta}, \hat\lambda_{\mathcal{M}_\lambda}, \hat\gamma_{\mathcal{M}^s_\gamma}; I) + log(n)\mathrm{tr}(\hat\Omega_\lambda \hat V_\lambda)\right], \qquad (6) \\
QIC^\gamma_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}^s_\lambda, \mathcal{M}_\gamma) &= \frac{1}{n}\left[-2Q_\gamma(\hat\beta_{\mathcal{M}^s_\beta}, \hat\lambda_{\mathcal{M}^s_\lambda}, \hat\gamma_{\mathcal{M}_\gamma}; I) + log(n)\mathrm{tr}(\hat\Omega_\gamma \hat V_\gamma)\right].
\end{aligned}
$$

where the explicit expressions of (6) are provided in Appendix C.

The three criteria in (6) are collectively named QIC_{JMVC}, so that model selection for three parameters β, λ and γ can then be conducted by minimizing those three criteria with respect to $\mathcal{M}_\beta$, $\mathcal{M}_\lambda$, and $\mathcal{M}_\gamma$, separately. It is worthwhile to point out that the reason why saturated estimators $\hat\beta_{\mathcal{M}^s_\beta}$, $\hat\lambda_{\mathcal{M}^s_\lambda}$ and $\hat\gamma_{\mathcal{M}^s_\gamma}$ are used is to avoid missing the important variables. In practice, however, if one has known the optimal model $\mathcal{M}^o_\beta$ for parameter β, one may also use $QIC^\lambda_{JMVC}(\mathcal{M}^o_\beta, \mathcal{M}_\lambda, \mathcal{M}^s_\gamma)$ to obtain $\mathcal{M}^o_\lambda$ and $QIC^\gamma_{JMVC}(\mathcal{M}^o_\beta, \mathcal{M}^s_\lambda, \mathcal{M}_\gamma)$ to select $\mathcal{M}^o_\gamma$. Thus the proposed criteria are flexible.

For computation, an efficient search strategy in spirit of Pan and Mackenzie (2003) is proposed as follows

$$
\begin{aligned}
\mathcal{M}^o_\beta &= \arg\min_{\mathcal{M}_\beta}\{QIC^\beta_{JMVC}(\mathcal{M}_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma)\}, \\
\mathcal{M}^o_\lambda &= \arg\min_{\mathcal{M}_\lambda}\{QIC^\lambda_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}_\lambda, \mathcal{M}^s_\gamma)\}, \\
\mathcal{M}^o_\gamma &= \arg\min_{\mathcal{M}_\gamma}\{QIC^\gamma_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}^s_\lambda, \mathcal{M}_\gamma)\}.
\end{aligned}
$$

It is clear that the number of minimization of this strategy is $2^p + 2^q + 2^d - 3$, which is computationally much less demanding than all subset selection number that is $(2^p -$

$1)(2^q-1)(2^d-1)$. The selection consistency of above algorithm is established by following theorem.

Theorem 4. *Under regularity conditions C1-C4 presented in supplementary material, as $n \to \infty$, we have*

$$Pr\{\min_{\mathcal{M}_\beta \in \mathrm{M}_\beta^+ \cup \mathrm{M}_\beta^-} QIC^\beta_{JMVC}(\mathcal{M}_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma) > QIC^\beta_{JMVC}(\mathcal{M}^o_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma)\ \} \to 1,$$

$$Pr\{\min_{\mathcal{M}_\lambda \in \mathrm{M}_\lambda^+ \cup \mathrm{M}_\lambda^-} QIC^\lambda_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}_\lambda, \mathcal{M}^s_\gamma) > QIC^\lambda_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}^o_\lambda, \mathcal{M}^s_\gamma)\ \} \to 1,$$

$$Pr\{\min_{\mathcal{M}_\gamma \in \mathrm{M}_\gamma^+ \cup \mathrm{M}_\gamma^-} QIC^\gamma_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}^s_\lambda, \mathcal{M}_\gamma) > QIC^\gamma_{JMVC}(\mathcal{M}^s_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^o_\gamma)\ \} \to 1.$$

The proof of Theorem 4 is presented in the supplementary materials. Zhang (2012) proposed a similar search strategy and proved the selection consistency. However, they assumed that the responses y_i must follow multivariate normal distribution, which is often violated in practice. Here we only assume the existence of the first four moments of y_i. Therefore our result is an improvement over their result and can be applied to broad range in practice.

5. Simulation Study

In this section, we present simulation results for JMVC estimators. The results confirm that (i) The JMVC estimators and corresponding estimation efficiency of mean parameter β are robust against misspecification of the working correlation structures. (ii) The proposed algorithm and sandwich formulas perform very well. (iii) The JMVC estimators of mean parameter β possess higher estimation efficiency than conventional GEE estimator. (iv) The model selection criterion QIC_{JMVC} performs well.

5.1. Simulation Setting

Without loss of generality, we focus on the case that σ^2_{ij} and ρ_{ijk} are actually not constants. We study the performance of JMVC estimators under different working correlation sturctures for R_{1i} and R_{2i} and the nuisance parameters u_1 and u_2. Denote CS(u_1) and AR(1)(u_2) by the matrices with Compound Symmetric structure and Order-1 Autoregressive structure with nuisance parameter u_1 and u_2, respectively. We study four cases shown in Table 1 with five sub-cases in each case. For example, Table 1 shows that

in the case I.(a) R_{1i} and R_{2i} are set as CS(0.3) and CS(0.3), respectively. We generate both normal data and normal-mixture data as two examples of continuous clustered data.

Table 1: Cases in simulation studies

Case	R_{1i}	R_{2i}	Case	R_{1i}	R_{2i}
I.(a)	CS(0.3)	CS(0.3)	III.(a)	CS(0.3)	AR(0.3)
I.(b)	CS(0.5)	CS(0.5)	III.(b)	CS(0.5)	AR(0.5)
I.(c)	CS(0.7)	CS(0.7)	III.(c)	CS(0.7)	AR(0.7)
I.(d)	CS(0.3)	CS(0.7)	III.(d)	CS(0.3)	AR(0.7)
I.(e)	CS(0.7)	CS(0.3)	III.(e)	CS(0.7)	AR(0.3)
II.(a)	AR(0.3)	AR(0.3)	IV.(a)	AR(0.3)	CS(0.3)
II.(b)	AR(0.5)	AR(0.5)	IV.(b)	AR(0.5)	CS(0.5)
II.(c)	AR(0.7)	AR(0.7)	IV.(c)	AR(0.7)	CS(0.7)
II.(d)	AR(0.3)	AR(0.7)	IV.(d)	AR(0.3)	CS(0.7)
II.(e)	AR(0.7)	AR(0.3)	IV.(e)	AR(0.7)	CS(0.3)

5.2. Normal Data

For each case in Table 5.1, we generate 1000 replicates, in which each with $n = 300$ clusters and each cluster has m_i observations with $m_i - 1 \sim Binomial(10, 0.7)$, resulting in different numbers of measurements for clusters. The replicates are generated from the model

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + e_{ij} \quad (i = 1, ..., n; j = 1, ..., m_i)$$
$$log(\sigma_{ij}^2) = \lambda_0 + z_{ij1}\lambda_1 + z_{ij2}\lambda_2 \quad \text{and} \quad f(\rho_{ijk}) = \gamma_0 + h_{ijk1}\gamma_1 + h_{ijk2}\gamma_2$$

where (x_{ij1}, x_{ij2}) are generated from $N_2(0, \text{CS}(0.5))$, $(z_{ij1}, z_{ij2}) = (x_{ij1}, x_{ij2})$ and $e_{ij} \sim N_{m_i}(0, \Sigma_i)$. For $h_{ijk} = (1, h_{ijk1}, h_{ijk2})^T$, since the generated correlation R_i should be positive definite, we present a proposition which provides a two-step algorithm below to generate such h_{ijk} in Appendix D.

Algorithm 2: Generating process of h_{ijk}

1: Generate (h_{ijk1}, h_{ijk2}) from $N_2(0, \text{CS}(0.3))$.

2: If $\|h_{ijk}\|_2 \leq \|\gamma\|_2^{-1} \min\{|f(-\frac{0.9}{m_i-1})|, |f(\frac{0.9}{m_i-1})|\}$, accept h_{ijk} as covariate of ρ_{ijk}. Otherwise return to step 1.

It is worthwhile to point out that we develop Algorithm 2 only for the purpose of generating the positive definite correlation matrices. In real application, statistical researchers often have information (e.g., from medical researchers) about covariates and have no need to generate covariates. However, when statistical researchers choose covariates based on such information, conditions presented in Algorithm 2 may not hold for h_{ijk} and the estimated correlation matrices $\hat{R}_i$ may not be positive definite. In such a situation, one could find a surrogate or calibration of $\hat{R}_i$. Specifically, one could find the surrogate by simply replacing the non-positive eigenvalues of $\hat{R}_i$ by its minimum positive eigenvalue or use the calibration techniques proposed by Huang et al. (2017) to find the positive definite calibration of $\hat{R}_i$.

From the generating process above, the generated data is unbalanced clustered data and there is no cluster-in order, thus MCD and HPC approaches do not applied to our simulation setting.

Due to space limit, only the simulation results for normal data of Case I and Case II are summarized in Table 2. The results of Case III and Case IV are presented in the supplementary materials. To evaluate the performance of sandwich formula (5), we also present the standard errors (in parentheses) and averaged standard deviations (in brackets) over 1000 parameter estimates for each case, where the standard deviations are estimated by the proposed sandwich formulas. In addition, we report estimation results of the mean parameters estimated by GEE using independent (*GEE. In*), Compound Symmetric (*GEE. CS*) and Order-1 Autoregressive (*GEE. AR(1)*) working correlation structures as competing methods, respectively.

In our simulation design, the average of the generated ρ_{ijk} over all observations is 0.0494, such small correlation is generated by setting small norm of the correlation parameter $\gamma = (\gamma_0, \gamma_1, \gamma_2)^T$, by which Algorithm 2 is easier to efficiently generate h_{ijk}. In Table 2, the small correlations lead to similar performance of *GEE. In*, *GEE. CS* and *GEE. AR(1)* in terms of efficiency, although the performance of *GEE. CS* and *GEE. AR(1)* is slightly superior than that of *GEE. In*.

It can be seen from Table 2 that both parameter estimates and their standard errors for regression parameters β are almost invariant against the working correlation structures and nuisance parameters of $R_{i1}(u_1)$ and $R_{i2}(u_2)$. Although there are slight

Table 2: Joint mean-variance-correlation estimation results of Case I and Case II (with sample standard errors in parentheses and sample estimated standard deviation in brackets) for normal data

		GEE			JMVC									
Parameter	True value	In	CS	AR(1)	I.(a)	I.(b)	I.(c)	I.(d)	I.(e)	II.(a)	II.(b)	II.(c)	II.(d)	II.(e)
β_0	1	0.9980	0.9978	0.9980	0.9969	0.9969	0.9969	0.9969	0.9969	0.9969	0.9969	0.9968	0.9969	0.9968
		(0.0834)	(0.0833)	(0.0834)	(0.0700)	(0.0700)	(0.0700)	(0.0700)	(0.0700)	(0.0700)	(0.0700)	(0.0702)	(0.0700)	(0.0700)
					[0.0679]	[0.0679]	[0.0678]	[0.0678]	[0.0679]	[0.0679]	[0.0679]	[0.0678]	[0.0678]	[0.0679]
β_1	-1	-1.0002	-1.0002	-1.0002	-1.0005	-1.0005	-1.0005	-1.0005	-1.0005	-1.0005	-1.0005	-1.0006	-1.0005	-1.0006
		(0.1093)	(0.1093)	(0.1093)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)	(0.0519)
					[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]
β_2	0.5	0.4998	0.4999	0.4998	0.5007	0.5007	0.5007	0.5006	0.5007	0.5007	0.5008	0.5008	0.5007	0.5008
		(0.1124)	(0.1119)	(0.1123)	(0.0508)	(0.0508)	(0.0508)	(0.0508)	(0.0508)	(0.0508)	(0.0509)	(0.0510)	(0.0508)	(0.0509)
					[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0491]	[0.0490]	[0.0490]	[0.0491]
λ_0	2	-	-	-	2.0082	2.0083	2.0083	2.0082	2.0083	2.0084	2.0085	2.0087	2.0084	2.0087
					(0.0310)	(0.0312)	(0.0313)	(0.0310)	(0.0313)	(0.0307)	(0.0317)	(0.0350)	(0.0307)	(0.0350)
					[0.0301]	[0.0302]	[0.0302]	[0.0301]	[0.0302]	[0.0302]	[0.0314]	[0.0349]	[0.0302]	[0.0349]
λ_1	1	-	-	-	1.0015	1.0015	1.0015	1.0015	1.0015	1.0012	1.0012	1.0012	1.0012	1.0012
					(0.0364)	(0.0370)	(0.0373)	(0.0364)	(0.0373)	(0.0367)	(0.0392)	(0.0415)	(0.0367)	(0.0415)
					[0.0337]	[0.0342]	[0.0345]	[0.0337]	[0.0345]	[0.0347]	[0.0373]	[0.0396]	[0.0347]	[0.0396]
λ_2	-1	-	-	-	-1.0013	-1.0012	-1.0012	-1.0013	-1.0012	-1.0015	-1.0015	-1.0014	-1.0015	-1.0014
					(0.0358)	(0.0363)	(0.0366)	(0.0358)	(0.0366)	(0.0358)	(0.0383)	(0.0406)	(0.0358)	(0.0406)
					[0.0337]	[0.0342]	[0.0344]	[0.0337]	[0.0344]	[0.0347]	[0.0373]	[0.0395]	[0.0347]	[0.0395]
γ_0	0.1	-	-	-	0.1003	0.1001	0.0996	0.0996	0.1003	0.1001	0.1001	0.1000	0.1000	0.1001
					(0.0290)	(0.0296)	(0.0323)	(0.0323)	(0.0290)	(0.0279)	(0.0282)	(0.0296)	(0.0295)	(0.0280)
					[0.0299]	[0.0324]	[0.0423]	[0.0423]	[0.0299]	[0.0286]	[0.0289]	[0.0311]	[0.0311]	[0.0286]
γ_1	-0.2	-	-	-	-0.1964	-0.1964	-0.1964	-0.1964	-0.1964	-0.1966	-0.1967	-0.1968	-0.1968	-0.1966
					(0.0223)	(0.0223)	(0.0224)	(0.0224)	(0.0223)	(0.0233)	(0.0252)	(0.0268)	(0.0263)	(0.0238)
					[0.0244]	[0.0244]	[0.0245]	[0.0245]	[0.0244]	[0.0255]	[0.0271]	[0.0284]	[0.0284]	[0.0255]
γ_2	0.15	-	-	-	0.1479	0.1479	0.1479	0.1479	0.1479	0.1480	0.1480	0.1480	0.1480	0.1480
					(0.0226)	(0.0227)	(0.0227)	(0.0227)	(0.0226)	(0.0236)	(0.0255)	(0.0270)	(0.0268)	(0.0238)
					[0.0237]	[0.0238]	[0.0238]	[0.0238]	[0.0237]	[0.0249]	[0.0265]	[0.0278]	[0.0278]	[0.0249]

perturbations in estimates of $\hat{\lambda}$ and $\hat{\gamma}$, they are consistent estimators. Besides, we can see clear changes of standard errors of $\hat{\lambda}$ and $\hat{\gamma}$, for example, looking at the standard errors of λ_1, it is 0.0364 in case I.(a), but in case II.(b), it becomes 0.0392. This is not unexpected since working correlations are used to approximate the true correlations and thus lead to some information loss. In addition, all sample estimated standard deviations match well with sample standard errors, indicating the sandwich formulas perform well. Compared with GEE approaches, the sample standard errors of $\hat{\beta}$ are uniformly smaller than those of *GEE. In*, *GEE. CS* and *GEE. AR(1)*. This is reasonable since *GEE. In*, *GEE. CS* and *GEE. AR(1)* do not identify the true covariance matrices, which are correctly identified by JMVC. Therefore, we conclude that that JMVC actually improves the estimation efficacy for the mean parameters.

In this simulation, there is no non-positive definite estimated correlation matrix appeared, which is because of the asymptotic consistency to the correlation matrix, see Theorem 3. In practice, however, the true correlation matrix may be non-positive definite. For such a situation, one can use the calibration technique proposed by Huang et al. (2017) to calibrate the estimated correlation matrix to ensure its the positive definiteness.

5.3. Normal-Mixture Data

In this simulation study we use the same setting as in section 5.2. We generate 1000 replicates from normal-mixture distributions

$$F_i = \pi N_{m_i}(\mu_i + a_i, \Sigma_i) + (1-\pi) N_{m_i}(\mu_i, \Sigma_i) \qquad (i = 1, ..., n)$$

where $\pi = 0.5$ is the mixing weight and $a_i = \frac{1}{2}\mu_i$ is the mean-shift parameter. For normal-mixture distribution F_i, the true expectation and variance are $\widetilde{\mu}_i = \mu_i + \pi a_i$ and $\widetilde{\Sigma}_i = \Sigma_i + \pi(1-\pi) a_i a_i'$, so that directly comparing the parameter estimators and the true values of parameters is not appropriate. Similar to Ye and Pan (2006), we use relative errors

$$\mathrm{err}(\hat{\mu}_i) = \frac{\|\hat{\mu}_i - \widetilde{\mu}_i\|}{\|\widetilde{\mu}_i\|} \qquad \mathrm{err}(\hat{\Sigma}_i) = \frac{\|\hat{\Sigma}_i - \widetilde{\Sigma}_i\|}{\|\widetilde{\Sigma}_i\|}$$

to measure the performance of our JMVC estimators, where $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are the estimated mean and covariance. The results are shown in Table 3.

Table 3: Joint mean-variance-correlation estimation results for averaged relative errors of normal-mixture data

	Case									
	I.(a)	I.(b)	I.(c)	I.(d)	I.(e)	II.(a)	II.(b)	II.(c)	II.(d)	II.(e)
err($\hat{\mu}$)	0.0601	0.0601	0.0601	0.0601	0.0601	0.0601	0.0602	0.0602	0.0602	0.0601
err($\hat{\Sigma}$)	0.1176	0.1180	0.1188	0.1183	0.1181	0.1179	0.1206	0.1238	0.1193	0.1225
	Case									
	III.(a)	III.(b)	III.(c)	III.(d)	III.(e)	IV.(a)	IV.(b)	IV.(c)	IV.(d)	IV.(e)
err($\hat{\mu}$)	0.0601	0.0602	0.0602	0.0602	0.0601	0.0601	0.0601	0.0601	0.0601	0.0601
err($\hat{\Sigma}$)	0.1179	0.1188	0.1198	0.1193	0.1184	0.1175	0.1199	0.1229	0.1182	0.1222

It can be seen that err($\hat{\mu}$) is almost robust against the change of working correlation structures and nuisance parameters of $R_{1i}(u_1)$ and $R_{2i}(u_2)$, whereas there is clear perturbation in err($\hat{\Sigma}$). This phenomenon coincides with the feature of JMVC estimators presented in Table 2 in the sense that the the parameter estimators and standard errors of β are robust, while there are perturbations in estimators and standard errors of $\hat{\lambda}$ and $\hat{\gamma}$. Overall, err($\hat{\mu}$) are negligible, whereas in some cases err($\hat{\Sigma}$) are relatively large but we consider this as acceptable since $\|\hat{\Sigma}_i - \widetilde{\Sigma}_i\|$ is for matrices with sizes $m_i \times m_i$.

5.4. Model Selection

Recall that the proposed selection criteria QIC_{JMVC} requires $2^p + 2^q + 2^d - 3$ times of minimization. For simplicity, we adopt the similar settings in section 5.2 to assess the performance of the proposed model selection criteria by setting $\beta = (1, -1, 0)^T$, $\lambda = (2, 1, 0)^T$ and $\lambda = (0.1, -0.2, 0)^T$.

Table 4 shows that the empirical percentage of the models which are incorrectly selected over 1000 replicates. Note that the performance of the correlation criteria QIC^{γ}_{JMVC} is best over all cases. A possible reason is that the generated correlations are small so that estimating the correlation parameters γ is easier. In our experience, the performance of the mean criteria QIC^{β}_{JMVC} is similar to that of QIC^{γ}_{JMVC} in real application.

Except Case II.(c), Case II.(e), Case III.(c) and Case III.(e), the mean criteria QIC^{β}_{JMVC} and the correlation criteria QIC^{γ}_{JMVC} are superior than the variance cri-

teria QIC^{λ}_{JMVC}, and the reason is that the variance is more difficult to estimate (Tang, 2011; Zhang et al., 2015). Therefore, in practice, we suggest to select the optimal model for mean and correlation first and then based on these two reasonable models to select the optimal model for variance. Overall, the percentage of incorrect selections for all three parameters is smaller than 0.165, which means a desired performance of our model selection criterion is obtained.

Table 4: Percentage of incorrectly selected models for normal data

	Case									
Criterion	I.(a)	I.(b)	I.(c)	I.(d)	I.(e)	II.(a)	II.(b)	II.(c)	II.(d)	II.(e)
QIC^{β}_{JMVC}	0.092	0.094	0.094	0.093	0.093	0.100	0.114	0.140	0.101	0.137
QIC^{λ}_{JMVC}	0.150	0.155	0.149	0.165	0.145	0.155	0.120	0.098	0.149	0.100
QIC^{γ}_{JMVC}	0.034	0.025	0.018	0.018	0.034	0.020	0.010	0.010	0.010	0.020
	Case									
Criterion	III.(a)	III.(b)	III.(c)	III.(d)	III.(e)	IV.(a)	IV.(b)	IV.(c)	IV.(d)	IV.(e)
QIC^{β}_{JMVC}	0.093	0.094	0.096	0.094	0.094	0.098	0.112	0.135	0.098	0.137
QIC^{λ}_{JMVC}	0.154	0.150	0.155	0.165	0.150	0.155	0.119	0.111	0.159	0.099
QIC^{γ}_{JMVC}	0.020	0.010	0.010	0.010	0.020	0.034	0.026	0.017	0.017	0.033

6. Analysis of COVID-19 data

Although the proposed approach focuses on cluster data with no need of cluster-order, in this section we analysis a COVID-19 data set, which is a longitudinal data with natural order in each cluster.

A global pandemic caused by Corona Virus Disease 2019 (COVID-19) leads to an urgent demand to understand this virus. One of the vital issues in COVID-19 research is to model the trajectory of the worldwide COVID-19 infection. A indicator that measures the severity of infection is the number of positive cases for each country. Here we use the data collected from website **Our World Data** (https://ourworldindata.org/) to model the trajectory of this indicator. The data consists of the number of positive cases of 95 countries in the world over a period from 21th September to 20th December. On 21th September, the number of positive cases varies from 4077 to 6976244 among these 95 countries. Daily counts of positive cases are recorded and are averaged every 7

Table 5: Table 5. Model selection results for global covid19 data

				p				
	2	3	4	5	6	7	8	9
$QIC^{\beta}_{JMVC}(p,13,4)$	13.5985	13.4528	13.3563	13.5835	13.5799	13.5810	13.5961	13.7013
				q				
	2	3	4	5	6	7	8	9
$QIC^{\lambda}_{JMVC}(4,q,4)$	45.6833	45.6940	46.2958	45.3660	47.3111	48.2647	48.2593	48.8596
				d				
	2	3	4	5	6	7	8	9
$QIC^{\gamma}_{JMVC}(13,13,d)$	135.5530	135.5493	135.5490	135.5492	135.5515	135.5521	135.5511	135.5521

days, leading to repeated measurements of 13 times. We treat the counts as continuous responses and we use three polynomials in time or time lag as covariates. That is, the covariates are of form

$$\begin{aligned} x_{ij} &= (1, t_{ij}, t_{ij}^2, ..., t_{ij}^{p-1})^T \\ z_{ij} &= (1, t_{ij}, t_{ij}^2, ..., t_{ij}^{q-1})^T \\ h_{ijk} &= (1, |t_{ij} - t_{ik}|, |t_{ij} - t_{ik}|^2, ..., |t_{ij} - t_{ik}|^{d-1})^T \end{aligned} \tag{7}$$

where t_{ij} is time at the jth measurements of the ith country.

(7) indicates that the model selection procedure reduces to select the best triple among all possible triples (p, q, d). We first use $QIC^{\gamma}_{JMVC}(13, 13, d)$ to select the appropriate d, where $QIC^{\gamma}_{JMVC}(p, q, d)$ is denoted by $QIC^{\gamma}_{JMVC}(\hat{\beta}^{\mathcal{M}^p_\beta}; \hat{\lambda}^{\mathcal{M}^q_\lambda}, \hat{\gamma}^{\mathcal{M}^d_\gamma})$ with $\mathcal{M}^p_\beta = \{1, ..., p\}$, $\mathcal{M}^q_\lambda = \{1, ..., q\}$, $\mathcal{M}^d_\gamma = \{1, ..., d\}$, respectively. The $QIC^{\beta}_{JMVC}(p, q, d)$ and $QIC^{\lambda}_{JMVC}(p, q, d)$ are defined similarly. We find that the optimal d is met at $d = 4$, and therefore we have information about optimal model for γ. After that, $QIC^{\beta}_{JMVC}(p, 13, 4)$ is used to select appropriate p, which is met at 4. Once having the optimal p and d, $QIC^{\lambda}_{JMVC}(4, q, 4)$ is used to select optimal q, which is met at 5. The selection results are shown in Table 5. From Table 5, the optimal triple for COVID-19 data clearly is $(4, 5, 4)$, indicating it suffices to model the mean by a three order polynomial, model the variance by a five order polynomial and model the correlation by a four order polynomial. We present our fitting results under optimal model in Figure 6.

Looking at Figure 6, three fitting curves are very close to the sample regressograms in terms of not only values and but also trajectory, indicating that the JMVC approach

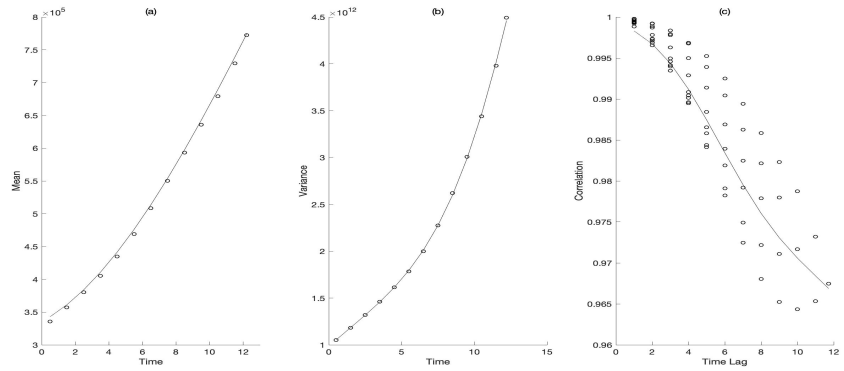

Figure 1: Analysis of COVID-19 data using best triple $(4, 5, 4)$: (a) shows mean versus time. (b) shows the variance versus time. (c) shows the correlation versus time lag. The scatters correspond to sample regressograms, the solid lines correspond to JMVC model fits.

produces a fairly satisfactory results. Note that there are clear patterns in the fitting curves, that is, there is a considerable upward trend of the number of positive cases across the world over this period. The rise of variance demonstrates the variability of worldwide positive cases goes up over time lag.

In this analysis, polynomials in time or time lag are used as covariates. This is due to the procedure that how one can apply selection criterion QIC_{JMVC} to real data can be presented clearly under such form of covariates. In other real applications, one may choose more general and iterpretable covarites to lend the advantage of JMVC model. For example, the stringency index, a indicator that measures the severity of epidemic prevention and control policy, may be a good choice as covariate for COVID-19 data. After having the estimated coefficients of more iterpretable covariates, the fitted JMVC model would have more a clear iterpration.

Another real data analysis of cattle data (Kenward and G., 1987) is presented in the supplementary materials, where the performance of JMVC and HPC (Zhang et al., 2015) is compared.

7. Discussions

In this paper, we propose a joint modelling approach for continuously clustered data. The mean, variance and correlation coefficients are modeled simultaneously. Rather than using any matrix decomposition techniques like Pan and Mackenzie (2003) and Zhang et al. (2015), the proposed approach directly models the original covariance matrix. Hence our approach permits clear interpretation of parameters. In addition, our approach does not require naturally ordered responses in each cluster, which is necessary in Pan and Mackenzie (2003) and Zhang et al. (2015). Since the covariance is correctly modeled, our approach improves estimation efficiency over the conventional GEE approach and the proposed mean parameter estimators and corresponding standard errors are robust against not only nuisance parameters but also structures of working correlation matrices. In addition the proposed model selection approach based on QIC is very flexible since the criteria in (6) can be used separately given information about optimal models.

As mentioned in section 3 and section 5.2, the estimated correlation matrix may be not positive definite, and one may use the calibration technique proposed by Huang et al. (2017) to ensure positive definiteness.

One future research interest is to extend the JMVC model to the context of semi-parametrical and nonparametrical statistics. We conjecture this is attainable by using the similar spline technique proposed in Leng et al. (2010). Another future interest is the JMVC model in the high dimensional setting, which means the dimension of covariate is allowed to diverge to infinity. Recently, Wang et al. (2012) proposed penalized generalized estimating equation (PGEE) approach to model high-dimensional clustered data, which is possible to incorporate into our approach. They used cross validation to select tuning parameter in PGEE, while we conjecture QIC_{JMVC} may be extended to high-dimensional setting. Finally, our approach focuses on continuous data. thus it is natural to investigate whether JMVC model may be extended to model discrete clustered data.

Acknowledgement

We would like to thank the editors and referees for their very constructive suggestions. This research is supported by the National Science Foundation of China (11871357) and

the Royal Society of the UK (R124683).

Appendix A. The calculation of $\mathbf{cov}(\boldsymbol{\epsilon}_{ij}^2, \boldsymbol{\epsilon}_{ik}^2)$ and $\mathbf{cov}(\boldsymbol{\delta}_{ijk}, \boldsymbol{\delta}_{ilm})$ under normal distribution

When y_i follows normal distribution $N_{m_i}(\mu_i, \Sigma_i)$, $\epsilon_{ij}/\sigma_{ij}$ follows normal distribution $N_1(0,1)$. Therefore by Wick's Theorem (Wick, 1950), we have

$$\mathrm{E}(\epsilon_{ij}^2\epsilon_{ik}^2) = \sigma_{ij}^2\sigma_{ik}^2\mathrm{E}\left[\left(\frac{\epsilon_{ij}}{\sigma_{ij}}\right)^2\left(\frac{\epsilon_{ik}}{\sigma_{ik}}\right)^2\right] = \mathrm{E}(\epsilon_{ij}^2)\mathrm{E}(\epsilon_{ik}^2)+2[\mathrm{E}(\epsilon_{ij}\epsilon_{ik})]^2 = \sigma_{ij}^2\sigma_{ik}^2+2\rho_{ijk}^2\sigma_{ij}^2\sigma_{ik}^2$$

Thus $\mathrm{cov}(\epsilon_{ij}, \epsilon_{ik}) = \mathrm{E}(\epsilon_{ij}^2\epsilon_{ik}^2) - \mathrm{E}(\epsilon_{ij}^2)\mathrm{E}(\epsilon_{ik})^2 = 2\rho_{ijk}^2\sigma_{ij}^2\sigma_{ik}^2$. We also have

$$\begin{aligned}\mathrm{E}\left(\frac{\epsilon_{ij}\epsilon_{ik}\epsilon_{il}\epsilon_{im}}{\sigma_{ij}\sigma_{ik}\sigma_{il}\sigma_{im}}\right) &= \mathrm{E}\left(\frac{\epsilon_{ij}\epsilon_{ik}}{\sigma_{ij}\sigma_{ik}}\right)\mathrm{E}\left(\frac{\epsilon_{il}\epsilon_{im}}{\sigma_{il}\sigma_{im}}\right)\\ &\quad+\mathrm{E}\left(\frac{\epsilon_{ij}\epsilon_{il}}{\sigma_{ij}\sigma_{il}}\right)\mathrm{E}\left(\frac{\epsilon_{ik}\epsilon_{im}}{\sigma_{ik}\sigma_{im}}\right)+\mathrm{E}\left(\frac{\epsilon_{ij}\epsilon_{im}}{\sigma_{ij}\sigma_{im}}\right)\mathrm{E}\left(\frac{\epsilon_{ik}\epsilon_{il}}{\sigma_{ik}\sigma_{il}}\right)\\ &= \rho_{ijk}\rho_{ilm}+\rho_{ijl}\rho_{ikm}+\rho_{ijm}\rho_{ikl}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{cov}(\delta_{ijk}, \delta_{ilm}) &= \mathrm{E}(\delta_{ijk}\delta_{ilm}) - \mathrm{E}(\delta_{ijk})\mathrm{E}(\delta_{ilm}) = \mathrm{E}\left(\frac{\epsilon_{ij}\epsilon_{ik}\epsilon_{il}\epsilon_{im}}{\sigma_{ij}\sigma_{ik}\sigma_{il}\sigma_{im}}\right) - \rho_{ijk}\rho_{ilm}\\ &= \rho_{ijl}\rho_{ikm}+\rho_{ijm}\rho_{ikl}\end{aligned}$$

Appendix B. The sandwich formula for covariances of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}$

Similar to insight of Liang and Zeger (1986), the estimated covariance of JMVC estimators $\hat{\lambda}_n$ and $\hat{\gamma}_n$ is given by

$$\begin{aligned}\mathrm{cov}(\hat{\lambda}_n) = &\left[\sum_{i=1}^n\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T\widetilde{W}_i^{-1}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)\right]^{-1}\left[\sum_{i=1}^n\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T\widetilde{W}_i^{-1}(\epsilon_i^2-\sigma_i^2)(\epsilon_i^2-\sigma_i^2)^T\widetilde{W}_i^{-1}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)\right]\\ &\times\left[\sum_{i=1}^n\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)^T\widetilde{W}_i^{-1}\left(\frac{\partial\sigma_i^2}{\partial\lambda}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}}\end{aligned}$$

and

$$\begin{aligned}\mathrm{cov}(\hat{\gamma}_n) = &\left[\sum_{i=1}^n\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T\widetilde{V}_i^{-1}\left(\frac{\partial\rho_i}{\partial\gamma}\right)\right]^{-1}\left[\sum_{i=1}^n\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T\widetilde{V}_i^{-1}(\delta_i-\rho_i)(\delta_i-\rho_i)^T\widetilde{V}_i^{-1}\left(\frac{\partial\rho_i}{\partial\gamma}\right)\right]\\ &\times\left[\sum_{i=1}^n\left(\frac{\partial\rho_i}{\partial\gamma}\right)^T\widetilde{V}_i^{-1}\left(\frac{\partial\rho_i}{\partial\gamma}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}}\end{aligned}$$

Appendix C. The explicit expression of model selection criteria

We now provide explicit expression for corresponding selection criteria in (6).

$$
\begin{aligned}
Q_\beta(\hat{\beta}_{\mathcal{M}_\beta}, \hat{\lambda}_{\mathcal{M}^s_\lambda}, \hat{\gamma}_{\mathcal{M}^s_\gamma}; I) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} Q_{1ij} \qquad Q_{1ij} = \int_{y_{ij}}^{\hat{\mu}_{ij}} \frac{y_{ij}-t}{\widetilde{\sigma}^2_{ij}}dt \\
\hat{\mu}_{ij} &= \mu_{ij}(x_{ij}^T\hat{\beta}_{\mathcal{M}_\beta}) \qquad \widetilde{\sigma}^2_{ij} = e^{z_{ij}^T\hat{\lambda}_{\mathcal{M}^s_\lambda}} \\
Q_\lambda(\hat{\beta}_{\mathcal{M}^s_\beta}, \hat{\lambda}_{\mathcal{M}_\lambda}, \hat{\gamma}_{\mathcal{M}^s_\gamma}; I) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} Q_{2ij} \qquad Q_{2ij} = \int_{\widetilde{\epsilon}^2_{ij}}^{\hat{\sigma}^2_{ij}} \frac{\widetilde{\epsilon}^2_{ij}-t}{2t^2}dt \\
\widetilde{\epsilon}^2_{ij} &= (y_{ij}-\mu_{ij}(x_{ij}^T\hat{\beta}_{\mathcal{M}^s_\beta}))^2 \qquad \hat{\sigma}^2_{ij} = e^{z_{ij}^T\hat{\lambda}_{\mathcal{M}_\lambda}}
\end{aligned}
\tag{C.1}
$$

$$
\begin{aligned}
Q_\gamma(\hat{\beta}_{\mathcal{M}^s_\beta}, \hat{\lambda}_{\mathcal{M}^s_\lambda}, \hat{\gamma}_{\mathcal{M}_\gamma}; I) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i-1}\sum_{k=j+1}^{m_i} Q_{3ijk} \qquad Q_{3ij} = \int_{\widetilde{\delta}_{ijk}}^{\hat{\rho}_{ijk}} \frac{\widetilde{\delta}_{ijk}-t}{1+t^2}dt \\
\widetilde{\delta}_{ijk} &= \frac{[y_{ij}-\mu(x_{ij}^T\hat{\beta}_{\mathcal{M}^s_\beta})][y_{ik}-\mu(x_{ik}^T\hat{\beta}_{\mathcal{M}^s_\beta})]}{\sigma_{ij}(z_{ij}^T\hat{\lambda}_{\mathcal{M}^s_\lambda})\sigma_{ik}(z_{ik}^T\hat{\lambda}_{\mathcal{M}^s_\lambda})} \qquad \hat{\rho}_{ijk} = f^{-1}(h_{ijk}^T\hat{\gamma}_{\mathcal{M}_\gamma})
\end{aligned}
\tag{C.2}
$$

$\hat{\Omega}_\beta\hat{V}_\beta$ is corresponding sandwich estimator under model $\mathcal{M} = (\mathcal{M}_\beta, \mathcal{M}^s_\lambda, \mathcal{M}^s_\gamma)$, where $\hat{\Omega}_\beta = -\partial Q_\beta/\partial\beta_{\mathcal{M}_\beta}\partial\beta^T_{\mathcal{M}_\beta}$ estimated at $(\hat{\beta}_{\mathcal{M}_\beta}, \hat{\lambda}_{\mathcal{M}^s_\lambda}, \hat{\gamma}_{\mathcal{M}^s_\gamma})$. $\hat{V}_\beta = \mathrm{cov}(\hat{\beta}_{\mathcal{M}_\beta})$ estimated at $(\hat{\beta}_{\mathcal{M}_\beta}, \hat{\lambda}_{\mathcal{M}^s_\lambda}, \hat{\gamma}_{\mathcal{M}^s_\gamma})$, $\hat{\Omega}_\lambda\hat{V}_\lambda$ and $\hat{\Omega}_\gamma\hat{V}_\gamma$ are similarly defined.

In (6), the normal distribution is used to approximate the distribution of y_i. Under normal distribution, $\mathrm{var}(\epsilon^2_{ij}) = 2\sigma^4_{ij} = 2[\mathrm{E}(\epsilon^2_{ij})]^2$, resulting in $2t^2$ in (C.1). And $1+t^2$ in (C.2) can be explained in same way.

Appendix D. Proposition to generate positive definite R_i

Proposition Appendix D.1. *For any $\alpha < 1$, under JMVC model, if $\|h_{ijk}\|_2 \le \|\gamma\|_2^{-1}\min\{|f(-\alpha/(m_i-1))|, |f(\alpha/(m_i-1))|\}$, then the correlation matrix R_i is positive definite.*

Proof: Since $\|h_{ijk}\| \le \|\gamma\|^{-1}\min\{|f(-\alpha/(m_i-1))|, |f(\alpha/(m_i-1))|\}$, by Cauchy-Schwarz inequality, we have

$$
|h_{ijk}^T\gamma| \le \|h_{ijk}\|\|\gamma\| \le \min\{|f(-\frac{\alpha}{m_i-1})|, |f(\frac{\alpha}{m_i-1})|\}
$$

Therefore by monotony of f

$$|\rho_{ijk}| = |f^{-1}(h_{ijk}^T \gamma)| \leq \frac{\alpha}{m_i - 1}$$

indicating that

$$\sum_{j \neq k} |\rho_{ijk}| \leq (m_i - 1)\frac{\alpha}{m_i - 1} \leq \alpha < 1$$

which means for each row of correlation matrix R_i, the sum of the absolute values of all non-diagonal elements is less than the diagonal element 1, since R_i is a symmetric matrix, it must be positive definite.

It is worthwhile to point out that this proposition holds for every finite d.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle, in: Selected papers of hirotugu akaike. Springer, pp. 199–213.

Crowder, M., 1995. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika 82, 407–410.

Daniels, M.J., Zhao, Y.D., 2003. Modelling the random effects covariance matrix in longitudinal data. Statistics in Medicine 22, 1631–1647.

Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S., 2002. Analysis of Longitudinal Data. Clarendon Press ;.

Fan, J., Huang, T., Li, R., 2007. Analysis of longitudinal data with semiparametric estimation of covariance function. Publications of the American Statistical Association 102, 632–641.

Huang, C., Farewell, D., Pan, J., 2017. A calibration method for non-positive definite covariance matrix in multivariate data analysis. Journal of Multivariate Analysis 157, 45–52.

Kenward, G., M., 1987. A method for comparing profiles of repeated measurements. Journal of the Royal Statistical Society. Series C, Applied statistics 36, 296–308.

Leng, C., Zhang, W., Pan, J., 2010. Semiparametric mean-covariance regression analysis for longitudinal data. Publications of the American Statistical Association 105, 181–193.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Pan, J., Mackenzie, G., 2003. On modeling mean-covariance structures in longitudinal studies. Biometrika 90, 239–244.

Pan, W., 2001. Akaikeš information criterion in generalized estimating equations 57, 120–125.

Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika , 677–690.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Tang, C.L.Y., 2011. Improving variance function estimation in semiparametric longitudinal data analysis. Canadian Journal of Statistics 39, 656–670.

Wang, L., Zhou, J., Qu, A., 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68, 353–360.

Wang, Y.G., Carey, V., 2003. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika 90, 29–41.

Wick, G.C., 1950. The evaluation of the collision matrix. Physical Review 80, 268–272.

Ye, H., Pan, J., 2006. Modelling of covariance structures in generalised estimating equations for longitudinal data. Biometrika 93, 927–941.

Zhang, W., 2012. A moving average cholesky factor model in covariance modelling for longitudinal data. Biometrika 99, p.141–150.

Zhang, W., Leng, C., Tang, C.Y., 2015. A joint modelling approach for longitudinal studies. Journal of the Royal Statal Society. Series B, Statistical methodology 77, 219–238.