A Joint Modeling Approach for Longitudinal Studies

Weiping Zhang, Chenlei Leng, and Cheng Yong Tang *

October 2, 2013

Abstract

In longitudinal studies, it is of fundamental importance to understand the dynamics in the mean function, variance function, and correlations of the repeated or clustered measurements. For modeling the covariance structure, Cholesky type decomposition based approaches are demonstrated effective. However, parsimonious approaches for directly revealing the correlation structure among longitudinal measurements remain less explored, and existing joint modeling approaches may encounter difficulty in interpreting the covariation structure. In this paper, we propose a novel joint mean-variance-correlation modeling approach for longitudinal studies. By applying hyperspherical coordinates, we obtain an unconstrained parametrization for the correlation matrix that automatically guarantees its positive definiteness, and develop a regression approach to model the correlation matrix of the longitudinal measurements by exploiting the parametrization. The proposed modeling framework is parsimonious, interpretable, and flexible for analyzing longitudinal data. Extensive data examples and simulations support the effectiveness of the proposed approach.

Some key words: Correlation matrix; Hyperspherical coordinates; Joint modeling; Longitudinal data analysis; Modified Cholesky decomposition.

*Zhang is with Department of Statistics and Finance, University of Science and Technology of China. Zhang's research is supported by NSF of China (No. 11271347, 11171321) (Email: zwp@ustc.edu.cn). Leng is with Department of Statistics, University of Warwick and Department of Statistics and Applied Probability, National University of Singapore (Email: C.Leng@warwick.ac.uk). Tang is with the Business School, University of Colorado Denver. Tang acknowledges research support from the Business School, University of Colorado Denver. (Email: chengyong.tang@ucdenver.edu). Corresponding author: Cheng Yong Tang. We thank the joint Editor, an associate editor, two referees and Prof. Paul Marriot for helpful comments.

1 Introduction

A longitudinal study involves observing the same variables repeatedly over a period of time, and is commonly encountered in psychology, social sciences, economics, and medical sciences. Since the collected observations of the same subject are not independent, it is fundamentally important to make effective use of the correlated measurements in analyzing data from longitudinal studies. Diggle et al. (2002) highlighted this issue and gave an excellent overview of various approaches to model this type of data sets.

Regression models on the mean and variance functions for understanding longitudinal data have been extensively studied in literature; see, for example, Liang and Zeger (1986), Lin and Carroll (2006), Fan et al. (2007), Fan and Wu (2008) and reference therein. Based on the mean-variance modeling framework, specifying the correlation structure by using, for example, the ARMA models with some index parameters has been explored; see, for example, Diggle et al. (2002), Fan et al. (2007), Fan and Wu (2008) and Qu et al. (2000). However, such statistical approaches do not permit more general forms of the correlation structure and cannot flexibly incorporate covariates that may help explain the covariations. To overcome this limitation, joint modeling for the mean and covariance becomes a poplar approach for longitudinal data analysis, and has received increasing interest recently; see, for example, Pourahmadi (1999, 2000, 2007), Pan and MacKenzie (2003), Ye and Pan (2006), Daniels and Pourahmadi (2009), Leng et al. (2010) and Zhang and Leng (2012). For parsimoniously modeling the covariations, a modified Cholesky decomposition was first applied by Pourahmadi (1999) to obtain unconstrained parametrization of the covariance matrix. An interesting aspect of such a decomposition is that the entries in this decomposition has autoregressive and log innovation interpretations. Pourahmadi (2007) pointed out that a decomposition studied by Chen and Dunson (2003) can be

understood as modeling certain moving average parameters and innovation variances. More recently, Zhang and Leng (2012) considered an alternative decomposition where the entries have moving average and log innovation interpretations; see also Yao and Li (2013) that uses the modified Cholesky decomposition in nonparametric regression for longitudinal data. A review of these approaches will be presented in a later section. Though demonstrated parsimonious and effective, these approaches can be viewed as indirectly modeling the variances and covariances of the longitudinal measurements. More specifically, due to the decompositions, the resulting variance functions of the aforementioned approaches cannot be directly interpreted as those of the repeated measurements. Moreover, the same interpretation issue also arises for the covariance and correlation structures when these approaches are applied. Therefore, for practical applications, additional effort and extra care are necessary for interpreting the resulting variance and covariance functions.

Not surprisingly, the development of a general regression approach to model the correlation structure is hindered by the requirements of a correlation matrix. Specifically, a correlation matrix has unity diagonal entries, and must be positive semi-definite with elements taking values between -1 and 1. Therefore, a regression approach based on a direct Cholesky type decomposition of the correlation matrix can hardly satisfy the requirements, and hence it encounters great difficulty in this scenario. In this paper, we propose a novel joint mean-variance-correlation modeling approach that targets directly the variances and correlations in the longitudinal data. Towards this end, we explore a new device for the correlation matrix by expressing it in hyperspherical coordinates using angles and trigonometric functions. Such a parametrization is very attractive because the resulting parameters are unconstrained on their support, and are directly interpretable with respect to the correlations. We then propose to construct a parsimonious regression model based on those angles and apply the maximum likelihood approach for parameter estimation and statisti-

cal inference. Our approach yields a parsimonious, interpretable, efficient and flexible framework for characterizing covariations in general correlated longitudinal data. Most importantly, such a parametrization and regression model are supported by data from real applications. For a balanced longitudinal data set in Section 3 where the empirical correlation matrix can be calculated from residuals, we can clearly see from Figure 3 that there exists some functional association between the angles and the time lag as a covariate. In addition, we also find that the proposed approach is preferred by comparing the Kullback-Leibler divergence and the information criteria of different approaches, in both simulations and data analysis. Along the line of exploring unconstrained parametrization, Daniels and Pourahmadi (2009) parametrized the correlation matrix via partial autocorrelations and their recursive relationships; but they studied neither efficient maximum likelihood estimation nor models for unbalanced data, popular in practical observational longitudinal studies. Compared to Daniels and Pourahmadi (2009), our method handles unbalanced longitudinal data more naturally.

The rest of this paper is organized as follows. Section 2 elaborates the new parametrization and its interpretations, the proposed joint modeling approach, the computational algorithm and its theoretical properties. We provide extensive numerical examples by applying our method to real data analysis including a balanced and an unbalanced data set, and conduct simulation comparisons in Section 3. The numerical results confirm the attractiveness of the new joint modeling approach. We conclude this paper by summarizing the main findings and outlining future research in Section 4. Technical proofs of the main results and additional interpretations of the new parametrization from the view points of geometric are given in the Supplementary Material of this paper.

2 Methodology

2.1 Longitudinal data modeling

Some notations are first introduced. We have generic longitudinal measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^{\mathrm{T}}$ $(i = 1, \ldots, n)$ collected from n subjects, observed at times $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})^{\mathrm{T}}$. Here we allow the number of repeated measurements m_i and time $\mathbf{t}_i$ to be subject specific, so that data sets observed at irregular time points and unbalanced longitudinal data can be analyzed using our framework. We denote by μ_{ij} and σ_{ij}^2 respectively the conditional mean and variance of y_{ij} given the covariate information at time t_{ij}. Modeling the quantities μ_{ij} and σ_{ij}^2 as functions of covariates by various methods is studied extensively in existing literature; see, for example, Diggle et al. (2002), Lin and Carroll (2006), Fan et al. (2007), Fan and Wu (2008), Wang (2003) and reference therein.

A crucial consideration in longitudinal data analysis is that the components of $\mathbf{y}_i$ are correlated, and it has been shown that incorporating the correlation is the key to efficiently utilizing data information in statistical inferences (Liang and Zeger, 1986; Diggle et al., 2002; Lin and Carroll, 2006). For simplicity and clarity in presentation, we assume hereinafter that $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$, $\boldsymbol{\Sigma}_i = \mathbf{D}_i \mathbf{R}_i \mathbf{D}_i$, $\mathbf{D}_i = \mathrm{diag}(\sigma_{i1}, \ldots, \sigma_{im_i})$, and $\mathbf{R}_i = (\rho_{ijk})_{j,k=1}^{m_i}$ is the correlation matrix of $\mathbf{y}_i$ with $\rho_{ijk} = \mathrm{corr}(y_{ij}, y_{ik})$ being the correlation between the jth and kth measurements of the ith subject. Here the distributional assumption can be relaxed and we may use estimating equations approaches (Liang and Zeger, 1986) instead of the likelihood approach for inferences, so that the general applicability of our method is not compromised in practice.

2.2 Existing approaches

In a class of modeling approaches, $\mathbf{R}_i$ is specified by some correlation matrix as a function of the observation time $\mathbf{t}_i$ – i.e., $\text{corr}(y_{ij}, y_{ik}) = \rho(t_{ij}, t_{ik}; \alpha)$ where $\rho(t_1, t_2, \alpha)$ is a positive definite function of t_1 and t_2, and is indexed by some unknown parameter α. We refer to Liang and Zeger (1986) as the seminal work of this class of approaches; see also Lin and Carroll (2006), Fan et al. (2007), Fan and Wu (2008) and reference therein for more recent development along this line of research. The limited choices of the positive definite function impose a severe restriction of the applicability of these approaches. To overcome the difficulty in specifying a model for $\mathbf{R}_i$, Qu et al. (2000) proposed to model $\mathbf{R}_i^{-1}$ by a weighted sum of a series of basis matrices – i.e., $\mathbf{R}_i^{-1} = \sum_{i=1}^{s} a_i \mathbf{M}_i$ for known matrices $\mathbf{M}_1, \ldots, \mathbf{M}_s$ and unknown constants $a_1, \ldots, a_s$. They then used the quadratic inference function approach for statistical inferences. Although this method often provides more efficient estimates for mean parameters, it is unclear how the covariations are affected by covariates.

In practice, it is desirable to study adaptive modeling approaches for $\mathbf{R}_i$, and to explore a broader range of information beyond $\mathbf{t}_i$ that might affect the covariations in $\mathbf{y}_i$. Pourahmadi (1999) proposed to parametrize the covariance matrix $\mathbf{\Sigma}_i$ by a modified Cholesky decomposition $\mathbf{P}_i \mathbf{\Sigma}_i \mathbf{P}_i^{\mathrm{T}} = \mathbf{\Gamma}_i^2$, where $\mathbf{P}_i = (p_{ijk})$ $(j, k = 1, \ldots, m_i)$ is a lower triangular matrix with 1's on its diagonal and $\mathbf{\Gamma}_i = \text{diag}\{\gamma_{i1}, ..., \gamma_{im_i}\}$ is a m_i-dimensional diagonal matrix. Immediately, we know that the lower triangular entries p_{ijk} are unconstrained. This decomposition is connected to the autoregressive model

$$y_{ij} - \mu_{ij} = -\sum_{k=1}^{j-1} p_{ijk}(y_{ik} - \mu_{ik}) + \epsilon_{ij}, \quad (i = 1, \ldots, n, j = 2, \ldots, m_i), \tag{1}$$

where ϵ_{ij} are independent innovations with innovation variance defined as $\text{var}(\epsilon_{ij}) = \gamma_{ij}^2$, and p_{ijk}'s are the so-called autoregressive coefficients, due to their similarity to the analogous components in time series analysis. Pourahmadi (1999) proposed to

link p_{ijk} to covariates via a regression model. By noting that $\mathbf{\Sigma}_i = \mathbf{Q}_i \mathbf{\Gamma}_i^2 \mathbf{Q}_i^{\mathrm{T}}$ where $\mathbf{Q}_i = (q_{ijk})$ $(j, k = 1, \ldots, m_i)$ is a lower triangular matrix with 1's on its diagonal and $\mathbf{\Gamma}_i$ is defined as above, Zhang and Leng (2012) characterized q_{ijk} as moving average parameters in the model $y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} q_{ijk}\epsilon_{ik} + \epsilon_{ij}$ $(i = 1, \ldots, n, j = 2, \ldots, m_i)$, and proposed to specify q_{ijk} as a parametrized function of covariates. Via a decomposition similar to that in Pourahmadi (1999), Chen and Dunson (2003) dealt with $\mathbf{\Sigma}_i = \mathbf{\Gamma}_i \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_i^{\mathrm{T}} \mathbf{\Gamma}_i$, where $\tilde{\mathbf{P}}_i = \mathbf{\Gamma}_i^{-1} \mathbf{Q}_i \mathbf{\Gamma}_i$. Pourahmadi (2007) interpreted the entries in this decomposition via a rescaled moving average model

$$\frac{y_{ij} - \mu_{ij}}{\gamma_{ij}} = \varepsilon_{ij} + \sum_{k=1}^{j-1} \tilde{p}_{ik} \varepsilon_{ik}, \quad (i = 1, \ldots, n, j = 2, \ldots, m_i),$$

where ε_{ij} are independent with $\mathrm{var}(\varepsilon_{ij}) = 1$. In this class of joint mean-covariance modeling approaches, however, components in $\mathbf{\Gamma}_i^2$ are not the conditional variances of the longitudinal response $\mathbf{y}_i$ given the covariates. To extract the variance information, one must transform the respective decompositions back to the original covariance matrix that gives nontrivial interpretations with respect to the covariates. Similarly, additional steps are also needed to study $\mathbf{R}_i$ as an objective of interest in practice for quantifying the correlations among the longitudinal measurements. Hence, extra effort and caution are required in practice to apply the aforementioned approaches for interpreting the features in the variance and covariations.

In a related work to ours, Daniels and Pourahmadi (2009) studied an alternative unconstrained parametrization by exploiting the partial autocorrelation matrix. They parametrized an $m \times m$ correlation matrix $\mathbf{R}$ by $\mathbf{\Pi} = (\pi_{ij})$ $(i, j = 1, \ldots, m)$ where $\pi_{ij} = \pi_{ji} = \mathrm{corr}(y_i, y_j | y_{i+1}, \ldots, y_{j-1})$ is the partial autocorrelation coefficient between y_i and y_j if $j > i + 1$, and otherwise $\pi_{ij} = \rho_{ij}$, the (i, j)th element of $\mathbf{R}$. In addition, the partial correlations connect to the correlations in $\mathbf{R}$ recursively via

$$\pi_{j(j+k)} = \frac{\rho_{j(j+k)} - \mathbf{r}_1^{\mathrm{T}}(j,k) \mathbf{R}_2(j,k)^{-1} \mathbf{r}_3(j,k)}{[1 - \mathbf{r}_1^{\mathrm{T}}(j,k) \mathbf{R}_2(j,k)^{-1} \mathbf{r}_1(j,k)]^{1/2} [1 - \mathbf{r}_3^{\mathrm{T}}(j,k) \mathbf{R}_2(j,k)^{-1} \mathbf{r}_3(j,k)]^{1/2}}$$

for $j = 1, \ldots, m - k; k = 1, \ldots, m - 1$, where $\mathbf{r}_1^{\mathrm{T}}(j,k) = (\rho_{j(j+1)}, \ldots, \rho_{j(j+k-1)})$,

$\mathbf{r}_3^{\mathrm{T}}(j,k) = (\rho_{(j+k)(j+1)}, \ldots, \rho_{(j+k)(j+k-1)})$ and $\mathbf{R}_2(j,k)$ is the correlation matrix corresponding to the components $j+1, \ldots, j+k-1$. A similar feature to our approach is that the elements π_{ij}'s in $\mathbf{\Pi}$ can vary freely in the interval $(-1,1)$ and take values in the entire real line after using Fisher's z transformation. Daniels and Pourahmadi (2009) focused on the Bayesian approach for the inference of parameters rather than studying the efficient maximum likelihood approach. Clearly, the mapping from correlation coefficients to partial autocorrelation coefficients is complicated, and the calculation of the Fisher information can be intractable for statistical inferences. It also remains less explored on how to apply this parametrization to unbalanced longitudinal data.

2.3 An unconstrained parametrization for correlation matrices and its interpretations

In this study, we propose to parametrize $\mathbf{R}_i$ via hyperspherical coordinates by the following decomposition:

$$\mathbf{R}_i = \mathbf{T}_i \mathbf{T}_i^{\mathrm{T}}, \tag{2}$$

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

$$\mathbf{T}_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{i21} & s_{i21} & 0 & \cdots & 0 \\ c_{i31} & c_{i32}s_{i31} & s_{i32}s_{i31} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{im_i1} & c_{im_i2}s_{im_i1} & c_{im_i3}s_{im_i2}s_{im_i1} & \cdots & \prod_{l=1}^{m_i-1} s_{im_il} \end{pmatrix} \tag{3}$$

and $c_{ijk} = \cos(\phi_{ijk})$ and $s_{ijk} = \sin(\phi_{ijk})$ are trigonometric functions of angles ϕ_{ijk}. In other words, the nonzero entries in the lower diagonal matrix $\mathbf{T}_i$ are given by $T_{i11} = 1$, $T_{ij1} = \cos(\phi_{ij1})$ for $j = 2, \ldots, m_i$, and

$$T_{ijk} = \begin{cases} \cos(\phi_{ijk}) \prod_{l=1}^{k-1} \sin(\phi_{ijl}), & 2 \le k < j \le m_i; \\ \prod_{l=1}^{k-1} \sin(\phi_{ijl}), & k = j; j = 2, \ldots, m_i. \end{cases} \tag{4}$$

Here the total number of angles ϕ_{ijk} $(1 \le k < j \le m_i)$ in (3) and (4) is $m_i(m_i-1)/2$, the same as that of the free parameters in an unconstrained correlation matrix. The

decomposition of $\mathbf{R}_i$ in (2) has several merits. First, columns of $\mathbf{T}_i^{\mathrm{T}}$ are unit vectors in $\mathbb{R}^{m_i}$, and the trigonometric expression (3) ensures the diagonal elements of $\mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}$ to be 1, and all other elements falling between -1 and 1. Second, $\mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}$ is always non-negative definite, satisfying the requirement of a correlation matrix. On the other hand, since $\mathbf{R}_i$ is a symmetric positive-semidefinite matrix, there always exists a lower triangular matrix $\mathbf{T}_i$ such that $\mathbf{R}_i = \mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}$. Following elementary algebra, we can take the angles in (3) and (4) as

$$\phi_{ijk} = \arccos\left(T_{ijk} / \prod_{l=1}^{k-1} \sin(\arccos(T_{ijl}))\right), 1 \leq k < j \leq m_i \tag{5}$$

where $\prod_1^0$ is taken as 1. Therefore, (5) can be viewed as a mapping from a general correlation matrix $\mathbf{R}_i$ to the angles. In addition, the mapping (5) is unique by restricting the range of the angles $\{\phi_{ijk}\}$ to be $[0, \pi)$ (Rapisarda et al., 2007). For unique model identification, we will take that all angles ϕ_{ijk} are in $[0, \pi]$ hereinafter. Hence a model for the angles $\{\phi_{ijk}\}$ in (3) is equivalent to a model for the correlation matrix. The most prominent advantage of (3) is that the angles $\{\phi_{ijk}\}$ as parameters are unconstrained in the range $[0, \pi)$. If needed, further transformation such as one involving arctan transform can give unconstrained parametrization in the entire real line $(-\infty, \infty)$. We observe in data analysis and simulation studies that it is sufficient to model the angles without further transformation. This facilitates a convenient and flexible modeling device for the correlation matrix. This parametrization was previously studied by Creal et al. (2011) in a different context for modeling correlations among multivariate financial time series.

By taking $\sum_1^0 = 0$ and noting that T_{ijk} in (4) can be recursively expressed by the correlations ρ_{ijk} as

$$T_{ijk} = \frac{\rho_{ijk} - \sum_{l=1}^{k-1} T_{ijl}T_{ikl}}{\sqrt{1 - \sum_{l=1}^{k-1} T_{ikl}^2}},$$

we observe from (5) that the angles can be expressed as functions of correlations. On

the other hand, from (2), ρ_{ijk} can be expressed by functions of the angles ϕ_{ijk} as

$$\rho_{ijk} = \cos(\phi_{ijk}) \prod_{l=1}^{k-1} \sin(\phi_{ijl}) \sin(\phi_{ikl}) + \sum_{l=1}^{k-1} \left[\cos(\phi_{ijl}) \cos(\phi_{ikl}) \prod_{t=1}^{l-1} \sin(\phi_{ijt}) \sin(\phi_{ikt}) \right] \quad (6)$$

for $1 \leq k < j \leq m_i$. We note that (6) establishes a hierarchic connection between the correlations and the angles. More specifically, ρ_{ijk} depends on ϕ_{ijk} only through $\cos(\phi_{ijk})$ upon given precedent angles ϕ_{ist} $(2 \leq s < j, 1 \leq t < k)$, or equivalently, upon given correlations ρ_{ist} $(2 \leq s < j, 1 \leq t < k)$. Such a hierarchic connection reflects the aforementioned intrinsic structural requirement for the correlation matrix. From (6), we see that $\partial \rho_{ijk} / \partial \phi_{ijk} = -\sin(\phi_{ijk}) \prod_{l=1}^{k-1} \sin(\phi_{ijl}) \sin(\phi_{ikl}) \leq 0$ because all angles are in $[0, \pi)$, implying that ρ_{ijk} is monotone decreasing in ϕ_{ijk}. Furthermore, because in practice the dependence between measurements in a longitudinal study generally decays with the time lag, small ρ_{ijk} is expected between measurements with large time lag. Since $\cos(\phi)$ is a monotone decreasing function on $[0, \pi)$, it is natural to expect from (6) that ϕ_{ijk} is increasing with the time lag between the jth and kth measurements. Our empirical data analysis also confirms this expectation; see Figures 3 and 6 in our numerical examples.

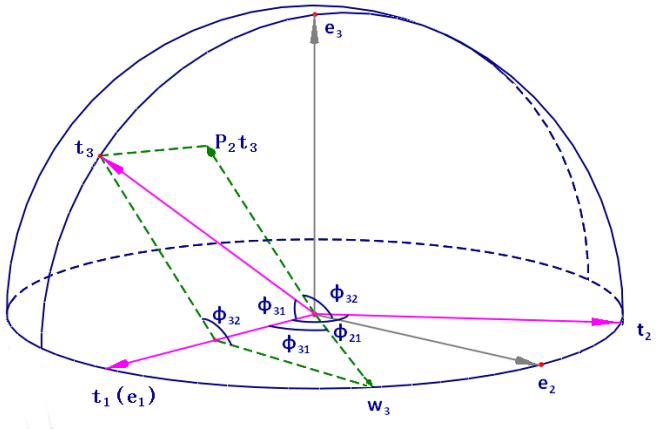

Figure 1: A 3-dimensional geometric representation of the correlations and the angles. Here, $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are the canonical basis, $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ are the three unit vectors in a unit ball where cosines of the pairwise angles are equal to the respective correlations between three longitudinal measurements, $\phi_{21}, \phi_{31}, \phi_{32}$ are the angles in the parametrization (2) and (4).

We now illustrate the geometric connection between the correlations and the angles via a graph when $m = 3$. In Figure 1, $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ are the three unit vectors in $\mathbf{T}$ on a 3-dimensional unit ball where cosines of the pairwise angles are equal to the respective correlations among three longitudinal measurements; $\phi_{21}, \phi_{31}, \phi_{32}$ are the angles in the parametrization (2) and (4). Clearly, ϕ_{21}, the angle between $\mathbf{t}_2$ and $\mathbf{t}_1$, and ϕ_{31}, the angle between $\mathbf{t}_3$ and $\mathbf{t}_1$, directly reflect the correlations between the first measurement and the other two. Once ϕ_{21} and ϕ_{31}, or equivalently, the correlations between the first and second measurements, and between the first and third measurements are specified, there is a one to one correspondence between ϕ_{32} and the correlation between the second and third measurement. Equivalently, there is a one-to-one correspondence between the three angles $\phi_{21}, \phi_{31}, \phi_{32}$ and the corresponding correlations among the three measurements. More detail on the general geometric interpretation of the new parametrization is available in the Supplementary Material of the paper.

2.4 The proposed approach

From the above discussion, $\mathbf{R}_i$ parametrized by angles ϕ_{ijk} ($i = 1, \ldots, n; 1 \leq k < j \leq m_i$) can be viewed as an equivalent expression of a correlation matrix in a hyperspherical coordinate system. Since $\mathbf{R}_i = \mathbf{T}_i \mathbf{T}_i^{\mathrm{T}}$ is guaranteed positive semidefinite and the angles in the parametrization (3) are unconstrained on $[0, \pi)$, we are free to characterize these angles via regression as functions of some covariates. In practice, such a rationale can be initially assessed by examining empirical variances and correlations from the observed longitudinal data. For a balanced longitudinal study such as the cattle example in Section 3, an initial version of the angles ϕ_{ijk} can be obtained from the empirical correlation matrix of the standardized residuals after a mean-variance model fitting. By examining the plot of those angles ϕ_{ijk} against the time lag in Figure 3, we clearly observe a curvature that supports some functional associations. From there, appropriate models can be used to describe such a curvature.

To illustrate the merits of the proposed parametrization more clearly, we now examine how the proposed parametrization behaves for two commonly used correlation structures.

Example 1. *For an $m \times m$ compound symmetry correlation matrix $\mathbf{R} = (1-\rho)\mathbf{I}_m + \rho\mathbf{J}_m$, where $\mathbf{I}_m$ is the m-dimensional identity matrix and $\mathbf{J}_m$ is a $m \times m$ matrix of 1s, the elements of matrix $\mathbf{T}$ can be seen as*

$$T_{11} = 1, T_{j1} = \rho, T_{jj} = (1-(j-1)\rho^2/(1+(j-2)\rho))^{1/2},\ j \geq 2;$$
$$T_{jk} = \rho\{1+(j-1)\rho\}^{-1}T_{kk},\ 2 \leq k < j \leq m.$$

Thus, the angles are specified by (5).

Example 2. *For an AR(1) correlation matrix $\mathbf{R} = (\rho^{|j-k|})_{j,k=1}^m$, the elements of $\mathbf{T}$ are found as $T_{j1} = \rho^{|j-1|}, j = 1, \ldots, m; T_{jk} = \rho^{|j-k|}/(1-\rho^2), 2 \leq k < j \leq m$. Then the angles are specified by (5), which are functions of ρ and time j and k.*

The above two examples show the equivalence of two parametrizations, one under the traditional and the other under the proposed parametrization. To examine our model more closely, we plot in Figure 2 angles ϕ_{jk}'s versus time lag $|j-k|$ for a compound symmetry and an AR(1) correlation structure with $\rho = 0.5$ respectively. Clearly, these scatter plots indicate some functional relationships approximately. More specifically, the compound symmetry structure can be explained by polynomials in time lag and a factor corresponding to the measurements ordering, while the AR(1) can be well explained by polynomials in time lag. Although not exact, a regression model constructed based on these angles represents a fairly good approximation to these commonly used correlations. Thus the proposed parametrization is flexible and adaptive for capturing the dynamics in these correlation structures.

Motivated by above considerations, we propose a joint regression model for the

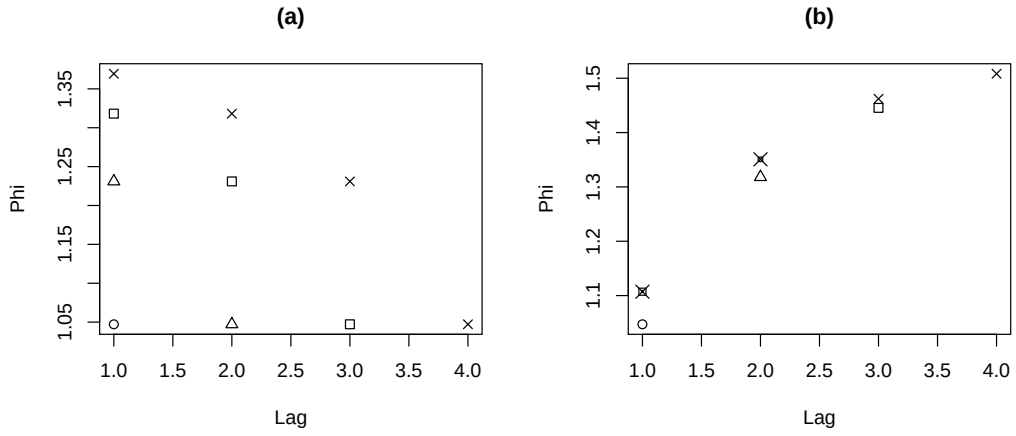

Figure 2: Plots of the angles ϕ versus lag for 5×5 (a) compound symmetry and (b) AR(1) correlation matrix with $\rho = 0.5$ respectively. Points are labeled by their row numbers in $\mathbf{T}_j$ where circle, triangle, square, and cross are corresponding to the second, third, fourth, and fifth rows.

mean, the variances, and the correlations as

$$g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \quad \log \sigma_{ij}^2 = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\lambda}, \quad \phi_{ijk} = \mathbf{w}_{ijk}^{\mathrm{T}}\boldsymbol{\gamma}, \tag{7}$$

where μ_{ij} and σ_{ij}^2 $(i = 1, \ldots, n, j = 1, \ldots, m_i)$ are respectively the conditional mean and variance functions for the jth measurement of the ith subject. As discussed earlier, for the ith subject, ϕ_{ijk} $(i = 1, \ldots, n; 1 \leq k < j \leq m_i)$ reflects the correlation between the jth measurement and the kth measurement once its correlations with the first $k - 1$ measurements are specified. We impose the regression model on the log-variance so that the variance function is automatically non-negative, and the support of $\boldsymbol{\lambda}$ is unconstrained. In this formulation, $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ and $\mathbf{w}_{ijk}$ are $p \times 1$, $q \times 1$ and $d \times 1$ vectors of generic covariates available, $g(\cdot)$ is a known link function, usually taken as an identity function as in linear models, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are unknown parameters for characterizing the mean, the variance and the correlation. The covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are specified in a way similar to those in heteroscedastic regression models. The covariate $\mathbf{w}_{ijk}$ should include time varying covariates that may depend on time t_{ij} and t_{ik} as it is used for capturing the correlation between the responses at these two times.

In practice, natural candidates for $\mathbf{w}_{ijk}$ include $(t_{ij}, t_{ik})^{\mathrm{T}}$ and its higher order terms, or a polynomial of the time lag $(t_{ik} - t_{ij})$ such that the resulting correlation is stationary (Ye and Pan, 2006). In practice, these covariates can be specified by graphical tools such as the technique we employed for analyzing the balanced cattle data in Section 3, or by model selection techniques where many potential covariates of interest are fitted initially and then selected. As for the range of ϕ_{ijk}, our experience from data analysis and simulations is that the estimated ϕ_{ijk}'s always fall in the range $[0, \pi)$. If its range is a concern, transformation such as arctan can be applied to ensure that ϕ_{ijk} falls in $[0, \pi)$. Remarkably, our framework generalizes easily to nonparametric and semiparametric models, although the focus of the paper is on parametric models as in (7).

To fit the joint model specified by (4) and (7) under the normality assumption, we note that

$$\frac{\partial T_{ijk}}{\partial \boldsymbol{\gamma}} = \begin{cases} T_{ijk}[-\mathbf{w}_{ijk} \tan(\phi_{ijk}) + \sum_{l=1}^{k-1} \mathbf{w}_{ijl} / \tan(\phi_{ijl})] & k < j \\ T_{ijk} \sum_{l=1}^{k-1} \mathbf{w}_{ijl} / \tan(\phi_{ijl}), & k = j \end{cases}, \tag{8}$$

where for simplicity, the notation $\sum_1^0$ is understood as 0 throughout this paper. By letting $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$, $\mathbf{D}_i = \mathrm{diag}(\sigma_{i1}, \ldots, \sigma_{im_i})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$, and noting that $\boldsymbol{\Sigma}_i = \mathbf{D}_i \mathbf{R}_i \mathbf{D}_i$, the minus twice log-likelihood, up to a constant, is given by

$$-2l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log |\mathbf{D}_i \mathbf{R}_i \mathbf{D}_i| + \sum_{i=1}^{n} \mathbf{r}_i^{\mathrm{T}} \mathbf{D}_i^{-1} \mathbf{R}_i^{-1} \mathbf{D}_i^{-1} \mathbf{r}_i, \tag{9}$$

where $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. We define $\boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i(\mathbf{X}_i \boldsymbol{\beta}) = \mathrm{diag}\{\dot{g}^{-1}(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}), \ldots, \dot{g}^{-1}(\mathbf{x}_{im_i}^{\mathrm{T}} \boldsymbol{\beta})\}$ where $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse link function $g^{-1}(\cdot)$ and we note that $\mu(\cdot) = g^{-1}(\cdot)$. We also define $\mathbf{b}_{ijk} = \sum_{l=k}^{j} \frac{\partial T_{ilk}}{\partial \boldsymbol{\gamma}} a_{ijl}$ with a_{ijl} being the (j, l) element of $\mathbf{T}_i^{-1}$, $\mathbf{h}_i = \mathrm{diag}\{\mathbf{R}_i^{-1} \mathbf{D}_i^{-1} \mathbf{r}_i \mathbf{r}_i' \mathbf{D}_i^{-1}\}$. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^{\mathrm{T}} = \mathbf{T}_i^{-1} \mathbf{D}_i^{-1} \mathbf{r}_i$, and thus $\epsilon_{i1}, \ldots, \epsilon_{im_i}$ are independent standard normal random variables, and denote by $\mathbf{1}_{m_i}$ the $m_i \times 1$ vector with elements 1. The following score equations based on the

likelihood can be obtained by direct calculations:

$$\begin{aligned}
\mathbf{U}_1(\boldsymbol{\beta};\boldsymbol{\gamma},\boldsymbol{\lambda}) &= \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \\
\mathbf{U}_2(\boldsymbol{\gamma};\boldsymbol{\beta},\boldsymbol{\lambda}) &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left\{ \frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} (\epsilon_{ij}^2 - 1) + \epsilon_{ij} \sum_{k=1}^{j-1} b_{ijk} \epsilon_{ik} \right\} = \mathbf{0}, \qquad (10) \\
\mathbf{U}_3(\boldsymbol{\lambda};\boldsymbol{\beta},\boldsymbol{\gamma}) &= \frac{1}{2} \sum_{i=1}^{n} \mathbf{Z}_i^{\mathrm{T}} (\mathbf{h}_i - \mathbf{1}_{m_i}) = \mathbf{0}.
\end{aligned}$$

We then estimate $\boldsymbol{\theta}$ by minimizing (9) via the iterative Newton-Raphson algorithm. Since the solutions satisfy equations (10), the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ can be sequentially solved one by one with other parameters kept fixed in the optimization. More specifically, we apply the following quasi-Fisher scoring algorithm:

1. Initialize the parameters as $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$. Set $k = 0$.

2. Compute $\boldsymbol{\Sigma}_i$ using $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\lambda}^{(k)}$. Update $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left[\mathbf{I}_{11}^{-1}(\boldsymbol{\theta}) \mathbf{U}_1(\boldsymbol{\beta};\boldsymbol{\gamma},\boldsymbol{\lambda}) \right] \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}. \qquad (11)$$

3. Given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, update $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ using

$$\begin{pmatrix} \boldsymbol{\gamma}^{(k+1)} \\ \boldsymbol{\lambda}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(k)} \\ \boldsymbol{\lambda}^{(k)} \end{pmatrix} + \left[\begin{pmatrix} \mathbf{I}_{22}(\boldsymbol{\theta}) & \mathbf{I}_{23}(\boldsymbol{\theta}) \\ \mathbf{I}_{32}(\boldsymbol{\theta}) & \mathbf{I}_{33}(\boldsymbol{\theta}) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{U}_2(\boldsymbol{\gamma};\boldsymbol{\beta},\boldsymbol{\lambda}) \\ \mathbf{U}_3(\boldsymbol{\lambda};\boldsymbol{\beta},\boldsymbol{\gamma}) \end{pmatrix} \right] \Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(k)},\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}}. \qquad (12)$$

4. Set $k \leftarrow k + 1$ and repeat Steps 2–3 until a pre-specified convergence criterion is met.

The expressions of $\mathbf{I}_{jk}(\boldsymbol{\theta}), j, k = 1, 2, 3$ are given in the next subsection. Note that this algorithm updates $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ together. This consideration is motivated by the asymptotic dependence of these two parameters, as can be seen from Theorem 1 in Section 2.5. Because the likelihood function is not a global convex function of the parameters on their support, it can only be guaranteed that the algorithm converges to a local optimum. To choose an appropriate initial value, we set $\boldsymbol{\Sigma}_i$'s as identity

matrices initially when solving for $\boldsymbol{\beta}$ in (11) by using the least square estimator as the initial value. Then we initiate $\boldsymbol{\gamma}$ in (12) by assuming $\mathbf{R}_i = \mathbf{I}_{m_i}$, the m_i dimensional identity matrix for the ith subject, and use the least square estimator based on the residuals to obtain an initial value of $\boldsymbol{\lambda}$. It is not difficult to see that these initial estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are $\sqrt{n}$-consistent. From the theoretical analysis in Theorem 1 in the next subsection and the proofs in the Appendix, the negative log-likelihood function is asymptotically convex around a small neighborhood of the true parameters. To ensure that the optimum is global, we may try multiple initial values for $\boldsymbol{\gamma}$. For our data analysis and simulation studies, the algorithm is quite stable with no multiple optima found and convergence was usually obtained within several iterations.

2.5 Theoretical properties

Let $\mathbf{I}(\boldsymbol{\theta}) = -E(\partial^2 l/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}})$ be the negative expected Hessian matrix. Our theoretical analysis assumes the following regularity conditions:

C1. The dimensions p, q and d of covariates $\mathbf{x}_{ij}, \mathbf{w}_{ijk}$ and $\mathbf{z}_{ij}$ are fixed; $n \to \infty$ and $\max_{1\le i\le n} m_i$ is bounded.

C2. The parameter space $\boldsymbol{\Theta}$ of $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ is a compact set in $\mathbb{R}^{p+q+d}$, and the true value $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}}, \boldsymbol{\lambda}_0^{\mathrm{T}})^{\mathrm{T}}$ is in the interior of $\boldsymbol{\Theta}$.

C3. As $n \to \infty$, $n^{-1}\mathbf{I}(\boldsymbol{\theta}_0)$ converges to a positive definite matrix $\mathcal{I}(\boldsymbol{\theta}_0)$.

Condition C1 is routinely made for longitudinal data from the practical perspective. Condition C2 is a conventional assumption for theoretical analysis of the maximum likelihood approach. Condition C3 is a natural requirement for the regression analysis in unbalanced longitudinal data modeling. We establish the following asymptotic results for the maximum likelihood estimator.

Theorem 1. *Under regularity conditions C1–C3, as $n \to \infty$, we have that: (a) the maximum likelihood estimator $(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\boldsymbol{\lambda}}^{\mathrm{T}})^{\mathrm{T}}$ is strongly consistent for the true*

value $(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}}, \boldsymbol{\lambda}_0^{\mathrm{T}})^{\mathrm{T}}$*; and (b)* $(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\boldsymbol{\lambda}}^{\mathrm{T}})^{\mathrm{T}}$ *is asymptotically normally distributed as* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\left[0, \{\mathcal{I}(\boldsymbol{\theta}_0)\}^{-1}\right]$*, where* $\mathcal{I}(\boldsymbol{\theta}_0)$ *is the Fisher information matrix defined in Condition C3.*

Following (10), it is shown in the Appendix that the block components of $\mathbf{I}(\boldsymbol{\theta})$ satisfy

$$\mathbf{I}_{11}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_i \mathbf{X}_i, \qquad \mathbf{I}_{12}(\boldsymbol{\theta}) = \mathbf{I}_{21}^{\mathrm{T}}(\boldsymbol{\theta}) = \mathbf{0}, \qquad \mathbf{I}_{13}(\boldsymbol{\theta}) = \mathbf{I}_{31}^{\mathrm{T}}(\boldsymbol{\theta}) = \mathbf{0},$$

$$\mathbf{I}_{23}(\boldsymbol{\theta}) = \mathbf{I}_{32}^{\mathrm{T}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[\frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} \mathbf{z}_{ij}^{\mathrm{T}} + \frac{1}{2} \sum_{k=1}^{j-1} \mathbf{b}_{ijk} \sum_{l=k}^{j} T_{ilk} a_{ijk} \mathbf{z}_{il}^{\mathrm{T}} \right],$$

$$\mathbf{I}_{22}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(2 \frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}} \frac{\partial \log T_{ijj}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} + \sum_{k=1}^{j-1} \mathbf{b}_{ijk} \mathbf{b}_{ijk}^{\mathrm{T}} \right),$$

$$\mathbf{I}_{33}(\boldsymbol{\theta}) = \frac{1}{4} \sum_{i=1}^{n} \mathbf{Z}_i^{\mathrm{T}} (\mathbf{I}_{m_i} + \mathbf{R}_i^{-1} \circ \mathbf{R}_i) \mathbf{Z}_i,$$

where $\circ$ denotes the Hadamard product. Since $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$ are consistent estimators for $\boldsymbol{\theta}_0$, the asymptotic covariance matrix $\mathcal{I}$ can be consistently estimated by a matrix whose block components are given by

$$\hat{\mathcal{I}}_{ij} = n^{-1} \mathbf{I}_{ij}(\hat{\boldsymbol{\theta}}), \quad (i, j = 1, 2, 3). \tag{13}$$

From Theorem 1, $\hat{\boldsymbol{\beta}}$ is asymptotically independent of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$. This is not surprising for statistical inferences of normally distributed data, because $\hat{\boldsymbol{\beta}}$ concerns the mean function, and $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$ are parameters of the covariations. From the generalized estimating equations point of view and (10), we also see that the optimal efficiency of estimating $\boldsymbol{\beta}$ is assured whenever $\boldsymbol{\Sigma}_i$'s or the models for σ_{ij}^2 and ϕ_{ijk} are correctly specified, a reminiscence of the results in generalized estimating equations by Liang and Zeger (1986). If the model for $\boldsymbol{\Sigma}_i$ is mis-specified, $\hat{\boldsymbol{\beta}}$ is still consistent and asymptotically normal by a result in Liang and Zeger (1986), although the asymptotic variance of $\hat{\boldsymbol{\beta}}$ would take a sandwich form. On the other hand, the two covariation parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$ are not asymptotically independent in general, unlike those

parameters in the modified Cholesky decomposition studied by Pourahmadi (1999) and Zhang and Leng (2012).

We now discuss a more general result that only requires the mean model in terms of $\boldsymbol{\beta}$ to be correctly specified. Recall that the Kullback-Leibler divergence $KL(f|f_0)$ between a true model with density f_0 with mean $\mu_0 = \mu(\mathbf{X}^{\mathrm{T}}\beta_0)$ and a working model $f_c \sim N_m(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma})$ is defined as

$$
\begin{aligned}
&KL(f_0|f_c) = E_{f_0}\log(\frac{f_0}{f_c}) = E_{f_0}\log(f_0) - E_{f_0}\log(f_c) \\
&= E_{f_0}\log(f_0) - E_{f_0}[-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu) - \frac{1}{2}\log|\Sigma| + \frac{m}{2}\log(2\pi)] \\
&= \frac{1}{2}\left[\mathrm{tr}(\Sigma_0\Sigma^{-1}) + (\mu_0-\mu)^T\Sigma^{-1}(\mu_0-\mu) + \log|\Sigma|\right] + E_{f_0}\log(f_0) - \frac{m}{2}\log(2\pi).
\end{aligned}
$$

We define a new population parameter vector $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_*^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}}, \boldsymbol{\lambda}_*^{\mathrm{T}})^{\mathrm{T}}$ to be the minimizer of $KL(f_0|f_c)$.

Corollary 1. *Under regularity conditions C1–C3 with $\boldsymbol{\theta}_0$ replaced by $\boldsymbol{\theta}_*$, as $n \to \infty$, we have that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is strongly consistent for $\boldsymbol{\theta}_*$; specifically if the mean model in (7) is correctly specified, $\boldsymbol{\beta}_* = \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of the true parameter vector $\boldsymbol{\beta}_0$.*

The first part of the corollary follows directly from Theorem 2.2 of White (1982). Furthermore, when the mean structure is correctly specified, we can see from the definition of $KL(f_0|f_c)$ that the minimizer $\boldsymbol{\beta}_*$ must equal to $\boldsymbol{\beta}_0$ if the mean function is correctly specified. Thus the consistency of $\hat{\boldsymbol{\beta}}$ neither relies on the normality assumption nor requires correct specification of the variance and correlation models in (7). This result is again similar to that in generalized estimating equations approaches (Liang and Zeger, 1986). As long as the mean model is correctly specified, the estimation of the covariance does not affect the consistency of the mean parameter, but only affects its efficiency.

3 Numerical Examples

3.1 Cattle data

We first apply our approach to a balanced longitudinal data set in Kenward (1987), where cattle were assigned randomly to two treatment groups A and B, and their weights were measured 11 times over a 133-day period. As in Pourahmadi (2000) and Pan and MacKenzie (2003), we focus on the 30 animals in group A using a saturated mean model with 11 parameters. Pan and MacKenzie (2003) found that it is suitable to apply three polynomials for characterizing the mean, the autoregressive coefficients and the log innovation variances in the model (1) based on the modified Cholesky decomposition. Using the Bayesian information criterion (BIC), they found the optimal triplet of the polynomial orders is (8,4,3).

Here we re-examine this data set with our joint modeling approach. For balanced longitudinal data, the corresponding angles ϕ_{ijk} of the empirical correlation matrix can be calculated using (5). By examining the angles versus the time lag between measurements in Figure 3, we see a clear curvature pattern that can be reasonably captured by a polynomial. Figure 3 also indicates a curvature pattern by examining the log sample variances versus the time of measurements.

Motivated by Figure 3, we model the angles in the proposed correlation parametrization by a polynomial of the time lag that defines $\mathbf{w}_{ijk}$. Following Pourahmadi (1999) and Pan and MacKenzie (2003), we apply two polynomials of time that define $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ for modeling the mean and the log-variances. The following BIC criterion is used to select the optimal model (Pan and MacKenzie, 2003; Zhang and Leng, 2012)

$$\text{BIC}(p, q, d) = -2\hat{l}_{max}/n + (p + q + d + 3)\log(n)/n, \tag{14}$$

where p, q, d are respectively the orders of three aforementioned polynomials, and $\hat{l}_{max}$ is the maximum of the corresponding log likelihood for the given orders. The BIC

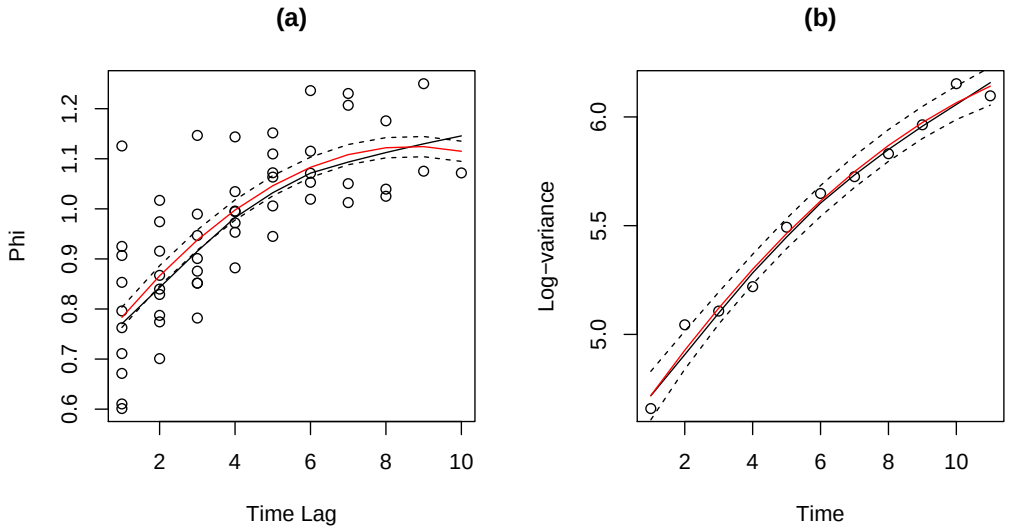

Figure 3: Sample regressograms and fitted curves for the cattle data: (a) shows the sample ϕ_{ijk} versus time lag; (b) shows the sample log-variance versus time. The solid black lines are fitted LOWESS lines, and the solid red lines are the fitted lines by the proposed model. Dashed curves represent asymptotic 95% pointwise confidence intervals.

criterion is known to be consistent in selecting the truth among candidate models (Shao, 1997). The optimal triplet using our approach is $(8, 2, 2)$ with a BIC value 52.03 and $\hat{l}_{max} = -755.0$. By comparing with the optimal triplet $(8, 4, 3)$ and $\hat{l}_{max} = -1045.40$ for Pan and MacKenzie (2003)'s best model, we clearly see that our method produces a larger likelihood and a smaller BIC value, indicating that the proposed parametrization better fits the data with a much more parsimonious model. It is interesting to see that it suffices to model the variances and the correlations via quadratic polynomials, yielding a simpler model than that from the approach in Pan and MacKenzie (2003). As discussed earlier, a likely reason for the improvement is that the proposed modeling approach is more flexible and adaptive so that it better fits the longitudinal data. We note that the fitted model for the angles implies a monotone relationship with the time lag, which is consistent with (6).

Table 1: Kenward's cattle data. Comparison of various models using the new joint modeling approach. *: The optimal triplet of the proposed method.

(p,q,d)	No.of parameters	$\hat{l}_{max}$	BIC
$(8,2,2)^*$	15	-755.00	52.03
(6,1,1)	11	-788.90	53.17
(3,3,3)	12	-800.27	54.71
(4,3,4)	14	-772.61	53.09
(7,2,2)	14	-761.67	52.37
(8,4,7)	22	-749.90	52.49
(9,3,1)	16	-756.92	52.28
(9,3,4)	19	-752.82	52.34
(9,5,8)	25	-748.29	52.72

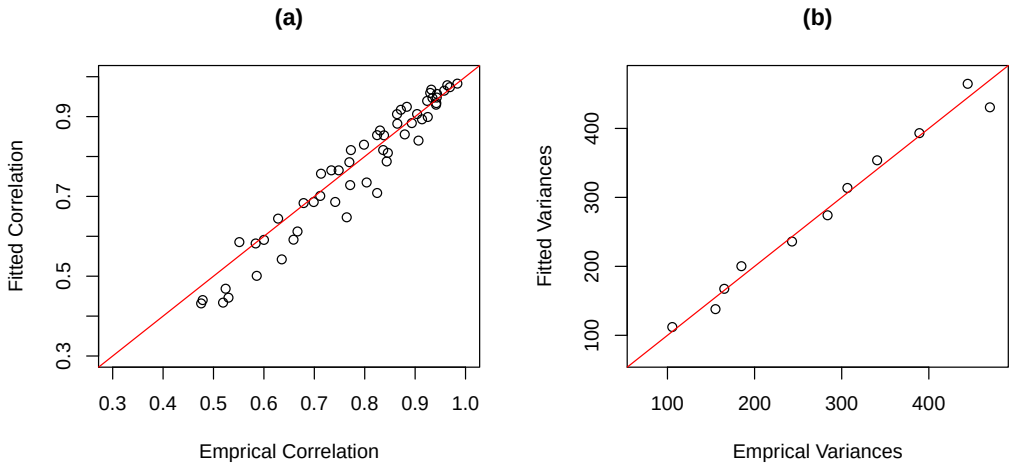

Figure 4: Plot of the empirical correlations and variances against the fitted correlations and variances; (a) correlations; and (b) variances.

The likelihood functions and the BIC values for a number of selected models are given in Table 1 for comparison purposes. Figure 4 depicts the fitted correlations and variances, versus the empirical correlations and variances respectively. Clearly, there is a close agreement between the sample quantities and the fitted quantities, implying that the proposed method produces a fairly good fit. To compare the accuracy of the empirical estimator and our estimator, we re-sample with respect to the subjects in the data set, and re-fit the optimal regression model. For each sample in the bootstrap, the empirical correlations and the estimated correlations are calculated.

Figure 5 shows that the standard deviations of the empirical correlations are generally larger that those of the fitted correlations by our approach. This is not surprising since we apply a model that captures the overall trend of these correlations as a function of time lag, and thus the noise in estimating the correlations is reduced by incorporating additional data information. We now report the time for model selection via all subset selection. Running on a Dell R910 2.00 GHz workstation, we specify the three polynomial orders from 1 to 11 for μ_{ij}, and from 1 to 10 for both ϕ_{ijk} and σ_{ij}. The user process time of the R code is about seven minutes for fitting $11 \times 10 \times 10$ models. Thus the computational time is manageable. We have also made use of a computational more efficient strategy for model selection as discussed in Zhang and Leng (2012) by fitting $11 + 10 \times 10$ models, thanks to the asymptotic orthogonality of the mean parameter and the variance parameters. The computational time has reduced about ten folds and the same optimal model is identified. Overall, we conclude that our approach is very effective for modeling the cattle data.

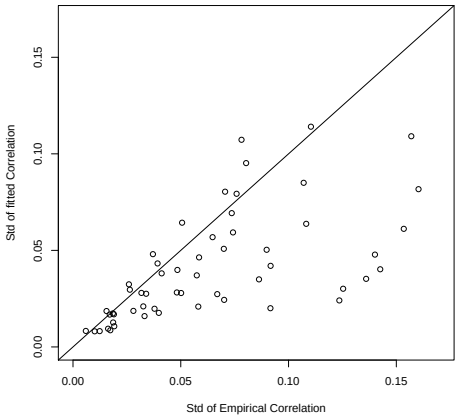

Figure 5: The standard deviations for the estimated correlations of the proposed approach and standard deviations of the empirical correlations. The standard deviations are calculated using a bootstrap approach that re-samples with respect to the subjects in Kenward's cattle data 100 times.

3.2 CD4 cell data

We apply the proposed joint modeling approach to an unbalanced data set, the CD4 cell study, previously analyzed by Zeger and Diggle (1994) and Ye and Pan (2006). The CD4 cell counts of 369 HIV-infected men with a total of 2,376 values were collected for this study, covering a period of approximately eight and a half years. This data set is observational and these counts were measured at different time for each individual. The number of measurements for each individual varies from 1 to 12 and the time points are not equally spaced. As in Zeger and Diggle (1994), to make the response variable more close to the Gaussian distribution, square roots of the CD4 counts are used. Clearly, this is a highly unbalanced data set and the method in Daniels and Pourahmadi (2009) developed mainly for balanced data sets is not applicable.

The objective of our analysis is to jointly model the mean, variance, and correlation structures. Using BIC for model selection, we find the optimal polynomials in our model, a degree eight polynomial in time for $\mathbf{x}_{ij}$ in the mean function, a linear function of time lag for $\mathbf{w}_{ijk}$ in the angles of the correlation model, and another linear function in time for $\mathbf{z}_{ij}$ in the log-variances. The optimal BIC value turns out to be 26.73, and $\hat{l}_{max} = -4892.72$. Figure 6 shows the fitted curves of the mean, angles in the correlation parametrization, and log-variances respectively. From Figure 6, we also observe the monotone increasing relationship of the fitted angles with the time lag. For comparison, we also apply the modified Cholesky decomposition approach in Pourahmadi (1999). The comparison is made in Table 2. We find that the best model of our approach is more parsimonious with a larger value in likelihood, and is more desirable in terms of the BIC value than the alternative. From Table 2, we can also find that our approach has larger values in likelihood and smaller BICs when compared at both the optimal models and models with the same complexity. Hence, we demonstrate the merits and wide applicability of our methods for general

unbalanced longitudinal data. When the highest polynomial orders of ten are used for selecting an optimal model for the three components in the joint modeling approach, the total CPU time for fitting 10^3 models is about 60 hours, and is reduced to about 6 hours if the computational thrifty strategy in Zhang and Leng (2012) is applied with the same optimal model identified.

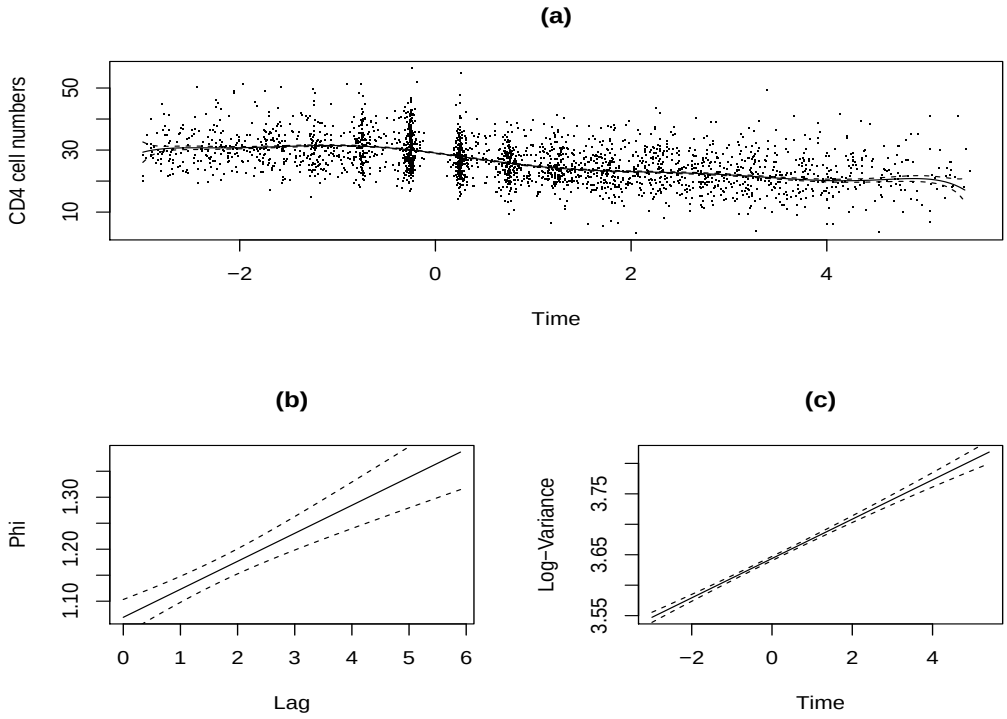

Figure 6: The CD4 cell data. The fitted curves of (a) the mean against time, (b) the angles against the time lag and (c) the log-variances against time. Dashed curves represent asymptotic 95% confidence intervals.

3.3 Simulation studies

In this section we investigate the finite sample performance of the proposed estimation and inference methods, and compare our method to the modified Cholesky decomposition approach. We conduct simulations in two studies.

Table 2: The CD4 data. Comparison of different models between our approach and the modified Cholesky decomposition approach (MCD) in Pourahmadi (2000) and Pan and MacKenzie (2003). *: The optimal triplet of the proposed method. **: The optimal triplet of the MCD method.

	No.of	Proposed		MCD	
(p,q,d)	parameters	$\hat{l}_{max}$	BIC	$\hat{l}_{max}$	BIC
$(8,1,1)^*$	13	-4892.72	26.72*	-5008.80	27.36
$(8,3,1)^{**}$	15	-4890.44	26.75	-4979.23	27.23**
(6,1,1)	11	-4902.17	26.75	-5018.47	27.38
(3,3,3)	12	-4919.52	26.85	-5006.18	27.33
(4,3,4)	14	-4902.10	26.80	-4995.51	27.30
(8,3,3)	17	-4886.36	26.76	-4974.70	27.24
(8,4,7)	22	-4881.76	26.81	-4971.74	27.30
(9,3,1)	17	-4888.34	26.75	-4983.73	27.27
(9,3,4)	19	-4881.30	26.76	-4974.15	27.26
(9,5,8)	26	-4877.16	26.84	-4968.30	27.33

Study 1. In the first study, data are generated from the proposed model and then the proposed approach is applied. This is to demonstrate the asymptotic properties in Section 2.5. The data sets are generated from the model

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + e_{ij},\ (i = 1, \ldots, n; \quad j = 1, \ldots, m_i),$$
$$\phi_{ijk} = \gamma_0 + w_{ijk1}\gamma_1 + w_{ijk2}\gamma_2,\ \log(\sigma_{ij}^2) = \lambda_0 + z_{ij1}\lambda_1 + z_{ij2}\lambda_2, \tag{15}$$

where $m_i - 1 \sim Binomial(6, 0.8)$, and then the measurement times t_{ij} are generated from the uniform(0,1) distribution. This setting results in different numbers of repeated measurements m_i for each subjects. The covariate $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})^{\mathrm{T}}$ is generated from a multivariate normal distribution with mean zero, marginal variance 1 and correlation 0.5. We take $\mathbf{z}_{ij} = \mathbf{x}_{ij}$, and $\mathbf{w}_{ijk} = \{1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2\}^{\mathrm{T}}$, following the setup in Leng et al. (2010). We generate 1000 data sets and consider sample sizes $n = 50$, 100, and 200 respectively.

Table 3 shows the accuracy of the estimated parameters in terms of their mean absolute biases (MAB) and standard deviations. All the biases are small especially when n is large. Additionally, to evaluate the inference procedure, we compare the sample standard deviation (SD) of 1,000 parameter estimates to the sample average of 1,000 standard errors (SE) using formula (13). The standard deviation (Std) of

1,000 standard errors is also reported. Table 3 shows that the SD and SE are quite close, especially for large n. This indicates that the standard error formula works well and demonstrates the validity of Theorem 1. Here we note that relatively higher level of variability is observed in estimating λ_0, the baseline level of the variance function, reflecting a fact that the variance function is generally harder to estimate in practice (Leng and Tang, 2011).

Table 3: Simulation results for Study 1 (all the results are multiplied by a factor 10^3).

	True	$n = 50$			$n = 100$			$n = 200$		
	value	MAB	SE_{Std}	SD	MAB	SE_{SD}	SD	MAB	SE_{SD}	SD
β_0	1.0	7.13	$7.83_{2.34}$	9.22	4.49	$5.13_{1.12}$	5.66	2.86	$3.50_{0.56}$	3.69
β_1	-0.5	1.67	$1.84_{0.58}$	2.18	1.05	$1.20_{0.28}$	1.32	0.67	$0.81_{0.14}$	0.86
β_2	0.5	1.01	$1.13_{0.35}$	1.32	0.64	$0.73_{0.17}$	0.81	0.41	$0.50_{0.09}$	0.53
γ_0	0.3	6.98	$7.31_{0.69}$	8.10	4.44	$5.15_{0.34}$	5.38	3.05	$3.64_{0.16}$	3.80
γ_1	-0.2	8.49	$8.48_{1.53}$	10.44	5.29	$5.89_{0.72}$	6.59	3.59	$4.14_{0.34}$	4.50
γ_2	0.3	9.46	$9.38_{1.37}$	11.14	5.83	$6.57_{0.64}$	7.17	3.92	$4.63_{0.30}$	4.92
λ_0	-0.5	110.42	$131.47_{5.15}$	139.54	74.01	$92.55_{2.58}$	92.92	51.19	$65.36_{1.26}$	65.29
λ_1	0.5	0.74	$0.65_{0.15}$	0.96	0.43	$0.46_{0.08}$	0.56	0.28	$0.32_{0.03}$	0.36
λ_2	-0.3	0.71	$0.66_{0.15}$	0.94	0.43	$0.46_{0.07}$	0.55	0.28	$0.32_{0.03}$	0.36

Study 2. In this study, we compare the proposed approach with the modified Cholesky decomposition (MCD) approach in Pourahmadi (1999) under different settings and for sample sizes $n = 50$ and 100. More specifically, we investigate three cases where the data models are either correctly or incorrectly specified respectively for our approach and the MCD approach. We fit the data using polynomials of different orders and always use a correct mean model for both approaches, as the key difference between these two approaches is how to model the covariance.

Case I. We take a similar model as in Study 1 to generate data sets with covariate $\mathbf{z}_{ij}$ taken as $\mathbf{z}_{ij} = (1, t_{ij}, t_{ij}^2)^{\text{T}}$. In this case, the covariance model for the MCD approach is mis-specified.

Case II. We generate data from model (1) following the MCD decomposition. A similar model structure as in Case I is implemented by changing ϕ_{ijk} in (15) to p_{ijk} in (1), and the variance function in (15) is used for the variance of ϵ_{ij} in (1). In this case, the variance function and correlation structure in our approach are mis-specified.

Case III. To compare these two methods when models are mis-specified for both approaches, we take the same mean model as in Case I with the marginal variance $\sigma^2(t) = 0.5e^t$ and ARMA(1,1) correlation structure $\text{corr}(\epsilon_t, \epsilon_s) = \gamma\rho^{|t-s|}$, for $t \neq s$. We consider $\gamma = 0.85$ and $\rho = 0.6$ corresponding to moderately correlated errors. In this case, both approaches use mis-specified models for the covariance, since this correlation structure does not exactly correspond to either decomposition. The best these two approaches can do is to capture some signals in this correlation with their respective model specifications.

Under each setting, polynomials of degrees q and d respectively are used for the angles in the correlation model and the log variances in our approach, and for the autoregressive coefficients and log innovations in the alternative MCD approach. The same orders of polynomials are applied for both approaches to make a fair comparison.

To compare these two methods, we define the following error measurements

$$\|\hat{\boldsymbol{\mu}}_d\| = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|,\ \|\hat{\boldsymbol{\Sigma}}_d\| = \frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_{0i}\|,\ \text{KL} = \frac{1}{n}\sum_{i=1}^{n} \text{KL}_i(f_{i1}|f_{i0}),$$

where KL is the Kullback-Leibler divergence between a fitted model $f_{i1} = N(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$ and the true model $f_{0i} = N(\boldsymbol{\mu}_{i0}, \boldsymbol{\Sigma}_{i0})$ for the ith subject. Table 4 shows the norms for the biases in the mean, the biases in the covariance and the KL divergences under different settings. In case I where the data are generated from our model, our approach performs substantially better than the alternative one in all comparing criteria, even for the saturated model ($q = 3, d = 3$) and a mis-specified reduced model ($q = 1, d = 2$). The KL divergences of the alternative approach are poor due to a lack of capability to capture the dynamics in this correlation structure. In Case II where the data are generated from the alternative MCD decomposition in (1), our approach still works reasonably well. The error measurements by our approach only inflate slightly compared to the alternative approach that fully exploits the model information. In case III where the data model is mis-specified for both approaches, ours works very promisingly. It is clear from Table 4 that our approach

substantially outperforms the alternative MCD approach in all measures. Though the KL divergences now take slightly larger values than those in Case I, they are much smaller than those of the alternative approach for all scenarios, indicating better fit to the truth. The simulations together with our data examples clearly demonstrate that the proposed approach is more adaptive and flexible for capturing the dynamics in the correlations of longitudinal data, even when the model is mis-specified.

Table 4: Comparison between the proposed method and Pourahmadi's modified Cholesky decomposition model (MCD).

	Proposed			MCD		
(q,d)	$\|\hat{\boldsymbol{\mu}}_d\|$	$\|\hat{\boldsymbol{\Sigma}}_d\|$	KL	$\|\hat{\boldsymbol{\mu}}_d\|$	$\|\hat{\boldsymbol{\Sigma}}_d\|$	KL
Case I, $n=50$						
(2,2)	$0.05_{0.04}$	$0.43_{0.35}$	$0.16_{0.11}$	$0.24_{0.17}$	$2.26_{0.23}$	$10.17_{0.58}$
(1,2)	$0.23_{0.18}$	$1.22_{0.39}$	$2.88_{0.46}$	$0.24_{0.17}$	$2.26_{0.23}$	$10.14_{0.57}$
(3,3)	$0.05_{0.04}$	$0.44_{0.36}$	$0.25_{0.17}$	$0.25_{0.17}$	$2.26_{0.23}$	$10.22_{0.60}$
Case I, $n=100$						
(2,2)	$0.03_{0.02}$	$0.29_{0.23}$	$0.06_{0.03}$	$0.18_{0.12}$	$2.24_{0.16}$	$10.04_{0.39}$
(1,2)	$0.16_{0.12}$	$1.27_{0.26}$	$2.69_{0.23}$	$0.18_{0.12}$	$2.25_{0.16}$	$10.04_{0.39}$
(3,3)	$0.03_{0.02}$	$0.30_{0.23}$	$0.08_{0.05}$	$0.18_{0.12}$	$2.25_{0.16}$	$10.05_{0.39}$
Case II, $n=50$						
(2,2)	$0.33_{0.15}$	$1.67_{0.36}$	$0.18_{0.07}$	$0.32_{0.15}$	$1.33_{0.54}$	$0.12_{0.06}$
(1,2)	$0.33_{0.15}$	$1.65_{0.36}$	$0.16_{0.06}$	$0.32_{0.15}$	$1.24_{0.54}$	$0.10_{0.06}$
(3,3)	$0.33_{0.15}$	$1.75_{0.37}$	$0.23_{0.21}$	$0.32_{0.16}$	$1.49_{0.53}$	$0.15_{0.08}$
Case II, $n=100$						
(2,2)	$0.23_{0.11}$	$1.45_{0.21}$	$0.11_{0.03}$	$0.23_{0.11}$	$0.92_{0.36}$	$0.06_{0.03}$
(1,2)	$0.23_{0.11}$	$1.44_{0.21}$	$0.11_{0.03}$	$0.23_{0.11}$	$0.87_{0.36}$	$0.05_{0.02}$
(3,3)	$0.23_{0.11}$	$1.49_{0.21}$	$0.13_{0.06}$	$0.23_{0.11}$	$1.03_{0.35}$	$0.07_{0.03}$
Case III, $n=50$						
(1,1)	$0.21_{0.12}$	$0.93_{0.30}$	$0.26_{0.05}$	$0.23_{0.13}$	$1.61_{0.19}$	$0.69_{0.07}$
(1,2)	$0.21_{0.12}$	$0.94_{0.30}$	$0.28_{0.06}$	$0.23_{0.13}$	$1.62_{0.19}$	$0.70_{0.08}$
(3,3)	$0.22_{0.12}$	$0.96_{0.29}$	$0.33_{0.10}$	$0.23_{0.13}$	$1.65_{0.18}$	$0.74_{0.10}$
Case III, $n=100$						
(1,1)	$0.15_{0.09}$	$0.86_{0.22}$	$0.21_{0.02}$	$0.16_{0.09}$	$1.57_{0.13}$	$0.64_{0.04}$
(1,2)	$0.15_{0.09}$	$0.87_{0.22}$	$0.22_{0.03}$	$0.16_{0.09}$	$1.58_{0.13}$	$0.64_{0.04}$
(3,3)	$0.15_{0.09}$	$0.88_{0.22}$	$0.24_{0.03}$	$0.17_{0.10}$	$1.59_{0.13}$	$0.66_{0.05}$

4 Discussion

We have proposed a novel joint approach for modeling the mean, the variance and the correlation in longitudinal data analysis. Our approach permits unconstrained parametrization, fast computation and easy interpretation of the parameters. Unlike previous approaches, this approach targets directly correlations and variances, and

provides the most general form of the covariance structure to our best knowledge.

Our decomposition opens many new avenues for future research. With unconstrained structures, we can model the mean, the variance and the correlations nonparametrically and semiparametrically. Along this line of research, for example, Wang et al. (2005), Fan et al. (2007), Fan and Wu (2008), and Yao and Li (2013) have developed methods for nonparametric approaches. The likelihood based estimation procedure permits regularization based model selection and it is interesting to study this further (Fan and Li, 2004; Bickel and Li, 2006). It may be interesting also to develop Bayesian inference procedures by eliciting appropriate priors. Finally, we assume that the longitudinal data follow multivariate normal distribution. It is worthwhile to further develop methods that are robust with respect to this assumption.

References

Bickel, P. and Li, B. (2006). Regularization in statistics (with discussion). *Test*, 15:271–344.

Chen, Z. and Dunson, D. (2003). Random effects selection in linear mixed models. *Biometrics*, 59:762–9.

Chiu, T. Y. M., Leonard, T., and Tsui, K. W. (1996). The matrix-logarithm covariance model. *Journal of American Statistical Association*, 91:198–210.

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.

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, 100:2352–2363.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of American Statistical Association*, 102:632–640.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of American Statistical Association*, 99:710–723.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of American Statistical Association*, 103:1520–1533.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, 36:296–308.

Leng, C. and Tang, C.Y. (2011). Improving variance function estimation in semiparametric longitudinal data analysis. *The Canadian Journal of Statistics*, 39:656–670.

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.

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

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of Royal Statistical Society, Series B*, 68:69–88.

Pan, J. and Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, 90:239–244.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86:677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87:425–35.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika*, 94:1006–1013.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving estimating equations using quadratic inference functions. *Biometrika*, 87:823–836.

Rapisarda, F., Brigo, D., and Mercurio, F. (2007). Parametrization correlations: a geometric interoretation. *IMA Journal of Management Mathematics*, 18:55–73.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90: 43–52.

Wang, N., Carroll, R. J., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100:147–157.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Yao, W. and Li, R. (2013). New local estimation procedure for a non-parametric regression function for longitudinal data. *Journal of the Royal Statistical Society, Series B*, 75:123–138.

Ye, H. and Pan, J. (2006). Modelling covariance structures in generalized estimating equations for longitudinal data. *Biometrika*, 93:927–941.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50:689–699.

Zhang, W. and Leng, C. (2012). A moving average cholesky factor model in covariance modeling for longitudinal data. *Biometrika,* 99:141-150.

Supplementary Material to "A Joint Modeling Approach for Longitudinal Studies"

Weiping Zhang, Chenlei Leng, and Cheng Yong Tang

This Supplementary Material contains more detail for geometrically interpreting the angles in the new parametrization (2) and (4), and technical proofs for the asymptotic properties of the estimation procedure.

1 More detail on the geometric interpretation

1.1 Geometric interpretation of the new parametrization

Now we describe a geometric view of a correlation matrix and the interpretation of the parametrization in (2) and (4). For clarity and simplicity, we consider a general m-dimensional correlation matrix $\mathbf{R} = (\rho_{jk})_{j,k=1}^m$. Let $\mathbf{T}$ be the corresponding lower triangular matrix defined in (2) such that $\mathbf{R} = \mathbf{T}\mathbf{T}^{\mathrm{T}}$, and denote by $\mathbf{T}^{\mathrm{T}} = (\mathbf{t}_1, \ldots, \mathbf{t}_m)$ where $\mathbf{t}_i$ $(i = 1, \ldots, m)$ are the column vectors in $\mathbf{T}^{\mathrm{T}}$. Therefore, we have $\rho_{jk} = \langle \mathbf{t}_j, \mathbf{t}_k \rangle$ $(j, k = 1, \ldots, m)$ where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. By observing that $\mathbf{t}_i$'s are all unit vectors, it is seen that the correlation ρ_{jk} is in fact the cosine of the angle between vectors $\mathbf{t}_j$ and $\mathbf{t}_k$. This suggests a geometric representation of a m-dimensional correlation matrix by m vectors $\mathbf{t}_1, \ldots, \mathbf{t}_m$ in a m-dimensional space with pairwise angles $\arccos(\rho_{jk})$. Indeed, such a geometric representation exists for any correlation matrix $\mathbf{R}$ and is unique by restricting the range of those angles to be $[0, \pi)$ (Rapisarda et al., 2007).

The geometric representation of a correlation matrix provides a natural way for interpreting the angles ϕ_{jk} $(1 \le k < j \le m)$ in (3) and (4) with respect to the correlations among longitudinal measurements. Let $\mathbf{e}_j = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots 0}_{m-j})^{\mathrm{T}}$ $(j = 1, \ldots, m)$ be the jth canonical basis of $\mathbb{R}^m$. Then, it is clear from (3) that $\mathbf{t}_1 = \mathbf{e}_1$ and $\langle \mathbf{t}_j, \mathbf{e}_1 \rangle = \cos(\phi_{j1}) = \rho_{j1}$ $(j = 2, \ldots, m)$. This implies that ϕ_{j1} is simply the angle

between $\mathbf{t}_j$ and $\mathbf{t}_1$ that reflects the correlation between the first measurement and the jth one. Further, let $\mathbb{P}_k = \mathrm{diag}(\underbrace{0,\ldots,0}_{k-1},\underbrace{1,\ldots,1}_{m-k+1})$ be the matrix such that $\mathbb{P}_k\mathbf{t}$ is the projection of the vector $\mathbf{t}$ into the subspace spanned by $\{\mathbf{e}_k,\ldots,\mathbf{e}_m\}$ for $k = 1\ldots,m$. For example, $\mathbb{P}_2\mathbf{t}_3 = (0,\cos(\phi_{32})\sin(\phi_{31}),\sin(\phi_{32})\sin(\phi_{31}),0,\ldots,0)^{\mathrm{T}}$ and therefore $\|\mathbb{P}_2\mathbf{t}_3\| = \sqrt{\langle\mathbb{P}_2\mathbf{t}_3,\mathbb{P}_2\mathbf{t}_3\rangle} = \sin(\phi_{31})$. Hence in this case $\langle\mathbb{P}_2\mathbf{t}_3,\mathbf{e}_2\rangle/\|\mathbb{P}_2\mathbf{t}_3\| = \cos(\phi_{32})$, implying that ϕ_{32} is the angle between $\mathbb{P}_2\mathbf{t}_3$ and $\mathbf{e}_2$. More generally, it can be shown analogously that ϕ_{jk} is the angle between $\mathbb{P}_k\mathbf{t}_j$ and $\mathbf{e}_k$ $(1 \le k < j \le m)$. On the other hand, since each $\mathbf{t}_j$ can be expressed as an orthogonal decomposition $\mathbf{t}_j = \sum_{i=1}^{m}\langle\mathbf{t}_j,\mathbf{e}_i\rangle\mathbf{e}_i$, the term $\mathbb{P}_k\mathbf{t}_j = \mathbf{t}_j - \sum_{i=1}^{k-1}\langle\mathbf{t}_j,\mathbf{e}_i\rangle\mathbf{e}_i$ denotes the remaining components in $\mathbf{t}_j$ by excluding the first $k-1$ coordinates in $\mathbb{R}^m$. By noting again the equivalence of a correlation and the cosine of an angle, we can see that ϕ_{jk} is the angle between the basis vector $\mathbf{e}_k$ and the remaining component in $\mathbf{t}_j$ after eliminating the contribution from the first $k-1$ coordinates.

By recalling the correspondence between vectors $\mathbf{t}_1,\ldots,\mathbf{t}_m$ in the geometric representation and the correlation matrix of m longitudinal measurements, we summarize the practical interpretations of the angles in our parametrization as follows. As seen from the hierarchic connection (6), for the jth measurement $(j = 2,\ldots,m)$, its correlations with the existing $j-1$ measurements are reflected by angles $\phi_{j1},\ldots,\phi_{j(j-1)}$. Specifically, $\cos(\phi_{j1})$ is its correlation with the first measurement. Once the correlations between the jth measurements and the first $k-1$ measurements $(2 \le k < j)$ are specified, the additional contribution from ϕ_{jk} properly reflects the correlation between the jth and kth measurements through the connection (6).

Another viewpoint of our parametrization is that a series of rotations of $\mathbf{t}_1$ can construct a geometric representation for the correlations between the three measurements. In particular, we first rotate $\mathbf{t}_1$ anti-clockwise with angle ϕ_{21} in the space spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$ to obtain $\mathbf{t}_2$. If we rotate $\mathbf{t}_1$ anti-clockwise with angle ϕ_{31} in the same space, we have an intermediate vector $\mathbf{w}_3$ in Figure 1. If we further rotate

$\mathbf{w}_3$ anti-clockwise with angle ϕ_{32} in the space parallel to the subspace spanned by $\mathbf{e}_2$ and $\mathbf{e}_3$, we obtain $\mathbf{t}_3$. Because the rotation of $\mathbf{w}_3$ is taken in a subspace orthogonal to $\mathbf{t}_1$ and $\mathbf{t}_2$, new contribution to $\mathbf{t}_3$ from the additional angle ϕ_{32} is orthogonal to $\mathbf{t}_1$ and $\mathbf{t}_2$ representing prior measurements. This process involves a series of Givens rotation and has been carefully studied by Rapisarda et al. (2007) in the context of understanding a correlation matrix $\mathbf{R}$ in a m-dimensional space. A brief description of this constructive interpretation is provided as follows.

1.2 Givens rotation for interpreting the angles in (2) and (4)

The connection between the correlations and the angles can be made from a constructive view of the vectors in the columns of $\mathbf{T}^{\mathrm{T}}$. The following shows how to construct $\mathbf{t}_1, \ldots, \mathbf{t}_m$ sequentially so that cosines of their pairwise angles are the correlations between the corresponding measurements. To this end, define a m-dimensional Givens rotation matrix as

$$\mathbf{G}(i,j;\phi) = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\phi) & \cdots & -\sin(\phi) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \sin(\phi) & \cdots & \cos(\phi) & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix},$$

where $\mathbf{G}(i,j;\phi)$ differs from the d-dimensional identity matrix in its (i,i), (i,j), (j,i), and (j,j) elements. For any vector $\mathbf{t} \in \mathbb{R}^m$, $\|\mathbf{G}(i,j;\phi)\mathbf{t}\| = \|\mathbf{t}\|$, i.e. the rotated vector and the original vector have the same length. Geometrically, the $\mathbf{G}(i,j;\phi)\mathbf{t}$ rotates the vector $\mathbf{t}$ anti-clockwise by angle ϕ in the subspace spanned by $\mathbf{e}_i$ and $\mathbf{e}_j$, while keeping all other components of $\mathbf{t}$ fixed. This property of the Givens rotation makes it an ideal device for constructing the m vectors $\mathbf{t}_1, \ldots, \mathbf{t}_m$ for representing a correlation $\mathbf{R} = (\rho_{ij})_{i,j=1}^m$. Following our earlier discussions in Subsection 2.3, there

is a one-to-one correspondence between $\mathbf{R}$ and angles ϕ_{ji} $(1 \leq i < j \leq m)$.

The procedure for constructing the vectors with those angles is as follows. Firstly, $\mathbf{t}_1$ is set as $\mathbf{e}_1$, and $\mathbf{t}_2$ is specified by rotating $\mathbf{e}_1$ in the subspace of $\mathbf{e}_1$ and $\mathbf{e}_2$ anti-clockwise with angle ϕ_{21} such that $\cos(\phi_{21}) = \rho_{12}$. With the Givens rotation matrix, we can write $\mathbf{t}_2 = \mathbf{G}(1, 2; \phi_{21})\mathbf{e}_1$, which is exactly the second column of $\mathbf{T}^{\mathrm{T}}$. Generally, the vector $\mathbf{t}_j$ with given $\mathbf{t}_1, \ldots, \mathbf{t}_{j-1}$ can be constructed by a total of $j-1$ anti-clockwise rotations from $\mathbf{e}_1$. The first rotation needs to satisfy that $\langle \mathbf{G}(1, 2; \phi_{j1})\mathbf{e}_1, \mathbf{t}_1 \rangle = \rho_{j1}$ and can be achieved easily by choosing the angle ϕ_{j1} properly. The second rotation needs to satisfy that $\langle \{\mathbf{G}(2, 3; \phi_{j2})\mathbf{G}(1, 2; \phi_{j1})\mathbf{e}_1\}, \mathbf{t}_2 \rangle = \rho_{j2}$. Because the Givens rotation only affects components in a two-dimensional subspace, we have $\langle \{\mathbf{G}(2, 3; \phi_{j2})\mathbf{G}(1, 2; \phi_{j1})\mathbf{e}_1\}, \mathbf{t}_1 \rangle = \rho_{j1}$ for any ϕ_{j2}. Therefore, ϕ_{j2} can be identified given a specified ϕ_{j1} in the first rotation. More generally, for the kth $(k < j)$ rotation,

$$\langle \prod_{i=1}^{k} \mathbf{G}(i, i+1; \phi_{ji})\mathbf{e}_1, \mathbf{t}_k \rangle = \rho_{jk}$$

is satisfied by appropriately choosing the angle ϕ_{jk} with given $\phi_{j1}, \ldots, \phi_{j(k-1)}$. By performing the rotation $j-1$ times following the above specification, the vector $\mathbf{t}_j$ is identified by

$$\mathbf{t}_j = \prod_{i=1}^{j-1} \mathbf{G}(i, i+1; \phi_{ji})\mathbf{e}_1.$$

The construction by the Givens rotation ensures that $\langle \mathbf{t}_j, \mathbf{t}_k \rangle = \rho_{jk}$ for all $k = 1, \ldots, j-1$. It can also be verified that $\mathbf{t}_j$ is exactly the jth column of $\mathbf{T}^{\mathrm{T}}$. By sequentially applying the above procedure for $j = 2, \ldots, m$, the vectors $\mathbf{t}_1, \ldots, \mathbf{t}_m$ can be obtained such that they form a geometric representation of the desirable correlation matrix $\mathbf{R}$.

2 Technical proofs

The Score and expectation of the Hessian

The computation of $\mathbf{U}_1(\boldsymbol{\beta};\boldsymbol{\gamma},\boldsymbol{\lambda})$ and $\mathbf{I}_{11}$ are trivial. Since $\boldsymbol{\Sigma}$ only depends on $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, it is easy to see that

$$\mathbf{I}_{12}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right) = -E\left[\sum_{i=1}^{n} X_i^{\mathrm{T}}\boldsymbol{\Delta}_i\frac{\partial\boldsymbol{\Sigma}_i^{-1}}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\{y_i - \boldsymbol{\mu}(\mathbf{X}_i\boldsymbol{\beta})\}\right] = 0.$$

Similarly $I_{13}(\boldsymbol{\theta}) = 0$. With $\mathbf{R}_i = \mathbf{T}_i\mathbf{T}_i^{\mathrm{T}}$, we have

$$-2l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\log\sigma_{ij}^2 + log(T_{ijj}^2) + \epsilon_{ij}^2).$$

Thus, the derivative of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\gamma}$ can be expressed by

$$\mathbf{U}_2(\boldsymbol{\gamma};\boldsymbol{\beta},\boldsymbol{\lambda}) = -\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}} + \frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\gamma}}\epsilon_{ij}\right). \tag{A.1}$$

As $\mathbf{T}_i\boldsymbol{\epsilon}_i = \mathbf{D}_i^{-1}\mathbf{r}_i$, it is easy to see that $\epsilon_{ij} = \frac{1}{T_{ijj}}(r_{ij}/\sigma_{ij} - \sum_{k=1}^{j-1}T_{ijk}\epsilon_{ik})$. Therefore, we have

$$\sum_{k=1}^{j} T_{ijk}\frac{\partial\epsilon_{ik}}{\partial\boldsymbol{\gamma}} = -\sum_{k=1}^{j}\epsilon_{ik}\frac{\partial T_{ijk}}{\partial\boldsymbol{\gamma}}, j = 1,\ldots,m_i,$$

or equivalently in matrix form, $\frac{\partial\boldsymbol{\epsilon}_i}{\partial\boldsymbol{\gamma}} = -(\boldsymbol{\epsilon}_i^{\mathrm{T}}\otimes\mathbf{I}_q)\frac{\partial\mathbf{T}_i^{\mathrm{T}}}{\partial\boldsymbol{\gamma}}\mathbf{T}_i^{-T}$. We then have

$$\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\gamma}} = -\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}}\epsilon_{ij} - \sum_{k=1}^{j-1}\mathbf{b}_{ijk}\epsilon_{ik}, \tag{A.2}$$

where $\mathbf{b}_{ijk} = \sum_{l=k}^{j}\frac{\partial T_{ilk}}{\partial\boldsymbol{\gamma}}a_{ijl}$ with a_{ijl} being the (j,l)th element of $\mathbf{T}_i^{-1}$. Combining (A.1) and (A.2) gives the second estimating equation in (10).

From (A.1) and (A.2), it is easy to know

$$\begin{aligned}
\mathbf{I}_{22}(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i}E\left(\frac{\partial^2 logT_{ijj}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}} + \frac{\partial^2\epsilon_{ij}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\epsilon_{ij} + \frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\gamma}}\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(2\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}}\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}^{\mathrm{T}}} + \sum_{k=1}^{j-1}\mathbf{b}_{ijk}\mathbf{b}_{ijk}^{\mathrm{T}}\right).
\end{aligned} \tag{A.3}$$

Similarly, we have

$$\begin{aligned}\mathbf{U}_3(\boldsymbol{\lambda};\boldsymbol{\beta},\boldsymbol{\gamma}) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\mathbf{z}_{ij}+2\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\lambda}}\epsilon_{ij}\right)\\ &= \frac{1}{2}\sum_{i=1}^{n}\left(-\sum_{j=1}^{m_i}\mathbf{z}_{ij}+\sum_{k=1}^{m_i}\mathbf{z}_{ik}\sum_{j=k}^{m_i}a_{ijk}\frac{r_{ik}}{\sigma_{ik}}\sum_{l=1}^{j}a_{ijl}\frac{r_{il}}{\sigma_{il}}\right) \qquad (A.4)\\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbf{Z}_i^{\mathrm{T}}(\mathbf{h}_i-\mathbf{1}_{m_i}),\end{aligned}$$

where the $d\times 1$ vector $\mathbf{h}_i=\mathrm{diag}\{\mathbf{R}_i^{-1}\mathbf{D}_i^{-1}\mathbf{r}_i\mathbf{r}_i'\mathbf{D}_i^{-1}\}$ and

$$\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\lambda}}=-\frac{1}{2}\sum_{k=1}^{j}\frac{r_{ik}}{\sigma_{ik}}a_{ijk}\mathbf{z}_{ik}, \qquad (A.5)$$

or in the matrix form $\partial\boldsymbol{\epsilon}_i/\partial\boldsymbol{\lambda}=-\frac{1}{2}\mathbf{Z}_i^{\mathrm{T}}\mathrm{diag}(\mathbf{D}_i^{-1}\mathbf{r}_i)\mathbf{T}_i^{-T}$.

From (A.1), (A.2) and (A.5) and the fact that $E\boldsymbol{\epsilon}_i\mathbf{r}_i^{\mathrm{T}}=\mathbf{T}_i^{\mathrm{T}}\mathbf{D}_i$, i.e.,

$$E\epsilon_{ij}\mathbf{r}_{ik}=\begin{cases}0, & k<j,\\ \sigma_{ik}T_{ikj}, & k\geq j;\end{cases}\quad j,k=1,\ldots,m_i,$$

we have

$$\begin{aligned}\mathbf{I}_{23}(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\lambda}^{\mathrm{T}}}\right)\\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m_i}E\left[\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}}\frac{\partial\epsilon_{ij}^2}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}+\sum_{k=1}^{j-1}\mathbf{b}_{ijk}\left(\epsilon_{ik}\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}+\frac{\partial\epsilon_{ik}}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}\epsilon_{ij}\right)\right]\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{\partial logT_{ijj}}{\partial\boldsymbol{\gamma}}\mathbf{z}_{ij}^{\mathrm{T}}+\frac{1}{2}\sum_{k=1}^{j-1}\mathbf{b}_{ijk}\sum_{l=k}^{j}T_{ilk}a_{ijk}\mathbf{z}_{il}^{\mathrm{T}}\right]. \qquad (A.6)\end{aligned}$$

Finally, from (A.4) we have

$$\begin{aligned}\mathbf{I}_{33}(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}}\right)=\sum_{i=1}^{n}\sum_{j=1}^{m_i}E\left(\frac{\partial^2\epsilon_{ij}}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}}\epsilon_{ij}+\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\lambda}}\frac{\partial\epsilon_{ij}}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}\right)\\ &= \frac{1}{4}\sum_{i=1}^{n}\mathbf{Z}_i^{\mathrm{T}}[\mathbf{I}_{m_i}+\mathbf{R}_i^{-1}\circ\mathbf{R}_i]\mathbf{Z}_i, \qquad (A.7)\end{aligned}$$

where $\mathbf{A}\circ\mathbf{B}$ denotes the Hadamard product of matrix $\mathbf{A}$ and $\mathbf{B}$.

Proof of Theorem 1. The proof is essentially the same as that of Theorem 1 in Pourahmadi (2000) and Theorem 1 and 2 in Chiu et al. (1996).

(a) Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ and $l_i = \log f_i(\mathbf{y}_i, \boldsymbol{\theta}), (i = 1, \ldots, n)$. Then ignoring the constant $\frac{1}{2} m_i \log(2\pi)$, we obtain that

$$l_i = -\frac{1}{2}\log(|\boldsymbol{\Sigma}_i|) - \frac{1}{2}\{\mathbf{y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta})\}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\{\mathbf{y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta})\}.$$

Thus the mean and the variance of l_i when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ are respectively

$$E_0(l_i) = -\frac{1}{2}\log(|\boldsymbol{\Sigma}_i|) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_{0i}) - \frac{1}{2}\{\mu(\mathbf{X}_i\boldsymbol{\beta}) - \mu(\mathbf{X}_i\boldsymbol{\beta}_0)\}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\{\mu(\mathbf{X}_i\boldsymbol{\beta}) - \mu(\mathbf{X}_i\boldsymbol{\beta}_0)\},$$
$$\mathrm{var}_0(l_i) = \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_{0i})^2 + 2\left\{\mu(\mathbf{X}_i\boldsymbol{\beta}) - \mu(\mathbf{X}_i\boldsymbol{\beta}_0)\right\}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_{0i}\boldsymbol{\Sigma}_i^{-1}\left\{\mu(\mathbf{X}_i\boldsymbol{\beta}) - \mu(\mathbf{X}_i\boldsymbol{\beta}_0)\right\}\right],$$

where $\boldsymbol{\Sigma}_i = \mathbf{D}_i\mathbf{R}_i\mathbf{D}_i^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_{0i} = \mathbf{D}_{0i}\mathbf{R}_{0i}\mathbf{D}_{0i}^{\mathrm{T}}$. It follows from the compactness of the parameter space and boundedness of the covariates that $\mathrm{var}_0(l_i) \leq K$, for all i where K is a constant. Therefore by Kolmogorov's strong law of large numbers, we have that

$$\frac{1}{n}\sum_{i=1}^{n} l_i - \frac{1}{n}\sum_{i=1}^{n} E_0(l_i) \to 0, \quad a.s.. \tag{A.8}$$

Notice that the above constant K is independent of $\boldsymbol{\theta}$ and it can be shown that $\frac{1}{n}\sum_{i=1}^{n} E_0(l_i(\boldsymbol{\theta}))$ is equicontinuous in $\boldsymbol{\theta}$, then following the proof of Theorem 1 in Chiu et al. (1996), it is easy to show the consistency of $\hat{\boldsymbol{\theta}}$.

The proof of (b) is essentially the same as that of Theorem 2 in Chiu et al. (1996). □