Robust estimation of mean and covariance for longitudinal data with dropouts

Guoyou Qin\textsuperscript{a} and Zhongyi Zhu\textsuperscript{b}\textsuperscript{*}

\textsuperscript{a}Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Fudan University, Shanghai 200032, People’s Republic of China; \textsuperscript{b}Department of Statistics, Fudan University, Shanghai 200433, People’s Republic of China

(Received 9 July 2014; accepted 12 December 2014)

In this paper, we study estimation of linear models in the framework of longitudinal data with dropouts. Under the assumptions that random errors follow an elliptical distribution and all the subjects share the same within-subject covariance matrix which does not depend on covariates, we develop a robust method for simultaneous estimation of mean and covariance. The proposed method is robust against outliers, and does not require to model the covariance and missing data process. Theoretical properties of the proposed estimator are established and simulation studies show its good performance. In the end, the proposed method is applied to a real data analysis for illustration.

Keywords: Cholesky decomposition; longitudinal data; mean–covariance; missing data; robustness

1. Introduction

Longitudinal data commonly arise in many research areas such as public health, medicine and social science, which is characterized by the dependence among the observations within the same subject. Different from independent data, within-subject covariance of the response plays an important role in the analysis of longitudinal data. For example, it is well known that ignoring the covariance may lead to loss of estimation efficiency for mean models [20,22,27]. In practice, the covariance itself is of great interest when the study aims to explore the association among the repeated responses. Therefore, simultaneous estimation of the mean and covariance has attracted substantial attention recently. For longitudinal studies with complete data, an incomplete list on this topic includes Pourahmadi [10,11], Ye and Pan [23], Leng \textit{et al.} [6], Mao \textit{et al.} [8], Zheng \textit{et al.} [26] and Yao and Li [22].

Compared with considerable literatures on simultaneous estimation of mean and covariance for longitudinal studies with complete data, the research for incomplete data is limited. Missing data are almost unavoidable due to various reasons during the period of longitudinal study which is designed to collect data for each subject at previously specified time points. Some subjects

\*Corresponding author. Email: zhuzy@fudan.edu.cn

© 2015 Taylor & Francis
may miss appointments or drop out before the end of follow-up. Excluding the missing data may produce biased estimates and thus result in invalid statistical inference. Hence, the classical statistical methods developed in the case of complete data may not work and some approaches for dealing with the missing data are proposed such as the inverse probability weighting (IPW) methods and imputation methods. There are large amounts of literatures on estimation of the mean for longitudinal data with missing values. For example, Robins et al. [17], Preisser et al. [12], Little and Rubin [7], Yi and He [24] and Chen and Zhou [2]. However, the research on estimation of both mean and covariance for longitudinal data with missing data is relatively limited. Recently, Garcia et al. [3] studied the estimation of mean and covariance for longitudinal data with response missing at random (MAR) based on simultaneously modeling the mean and covariance under the normality assumption of the random errors.

Generalized estimating equation (GEE) methods or likelihood methods are generally adopted for the estimation of the joint mean and covariance models. However, it is well known that both GEE and likelihood methods are sensitive to the outliers in the data. Recently, Zheng et al. [26] proposed a robust GEE method for the estimation of the joint mean–covariance model in the case of complete data. Qin et al. [14] developed a robust IPW–GEE method for the estimation of the mean for longitudinal data with dropouts. To the best of our knowledge, the work on robust estimation of both mean and covariance has not been discussed.

In this paper, we aim at the development of robust estimation of both mean and covariance for longitudinal data with dropouts. Under assumptions that random errors follow an elliptical distribution and all the subjects share the same within-subject covariance matrix which does not depend on covariates, we propose a robust approach to achieve this goal. The advantages of the proposed method can be summarized as follows. First, the proposed estimator is robust against outliers in the data. Second, unlike the common approach to model the covariance such as Garcia et al. [3], our method does not need to specify the covariance structure or model the parameters of the covariance. Thus, the issue of model misspecification can be avoided. Third, specification of the model for the missing data process is not required. In brief, the proposed estimator for the mean and covariance is robust against outliers and does not require to model the covariance and missing data process.

The paper is organized as follows. We will describe the model and develop the proposed method in Sections 2 and 3, respectively. Theoretical properties of the proposed estimator are presented in Section 4 and simulation studies are conducted in Section 5. In the end, the proposed method is applied to analyze a real data set.

2. Models

Suppose that a longitudinal study involves $n$ subjects with $m$ observations over time for each subject. A linear regression model is considered to model the relationship between the response $Y_{ij}$ and the covariates $X_{ij}$ as

$$Y_{ij} = X_{ij}^T \beta_0 + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

where $\beta_0$ is a $p$-dimensional vector of regression coefficients associated with $X_{ij}$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T$ independently follow continuous elliptical distribution with mean $0_{m \times 1}$ and covariance matrix $\Sigma_0$. The whole class of continuous elliptical distributions is a very rich family, which can be used in many statistical procedures instead of the Gaussian distributions [4]. At time point $j$, the complete data for the $i$th subject are $(Y_{ij}, X_{ij})$, where $Y_{ij}$ is a response variable and may be missing, and $X_{ij}$ is a $p$-dimensional vector of covariates. Denote $Y_i = (Y_{i1}, \ldots, Y_{im})^T$ and $X_i = (X_{i1}, \ldots, X_{im})^T$. In this paper, we focus on robust estimation of both $\beta_0$ and $\Sigma_0$ in the case of MAR.
Let \( R_{ij} = 1 \) if \( Y_{ij} \) is observed; \( R_{ij} = 0 \), otherwise. We here consider the case of response MAR. The MAR mechanism is commonly assumed in the literatures on missing data since this mechanism is usually in accordance with most practical issues \cite{9}. The response MAR means that the distribution of the missing data indicators \( R_{i} = (R_{i1}, \ldots, R_{jm})^T \) conditional on \( Y_{i} \) and \( X_{i} \) only depends on the observed data. Denote \( \hat{R}_{ij} = \{R_{i1}, \ldots, R_{j-1}\}^T \) as the history of missing data indicators at time point \( j \), and let \( Y_{ij}'' \) denote the observed components of \( Y_{i} \). Then, the MAR mechanism considered here can be clearly expressed as \( P(R_{ij} = 1 | \hat{R}_{ij}, Y_{i}, X_{i}) = P(R_{ij} = 1 | \hat{R}_{ij}, Y_{ij}'' , X_{i}) \). Furthermore, we here consider the monotone missing pattern, that is, \( R_{ij} = 0 \) indicates \( R_{ik} = 0 \) for all \( k > j \). We assume \( R_{i1} = 1 \) for each subject.

### 3. Proposed method

In this section, we develop a robust method for simultaneous estimation of the mean and covariance.

According to Cholesky decomposition, there exists a lower triangle matrix \( C \) with 1s on the main diagonal such that

\[
\text{cov}(C_{ij}) = C \Sigma_0 C^T = B, \tag{2}
\]

where \( B \) is a diagonal matrix. Denote \( e_i = (e_{i1}, \ldots, e_{im})^T = C e_i \) and then we have \( e_{i1} = e_{i1} \) and \( e_{ij} = c_{j1} e_{i1} + c_{j2} e_{i2} + \cdots + c_{jj-1} e_{jj-1} + e_{ij}, \ i = 1, \ldots, n, \ j = 2, \ldots, m \), where \( c_{jk} \) is the negative of the \((j, k)\)-component of the matrix \( C \). Denote \( B = \text{diag}(b_{11}^2, \ldots, b_{mm}^2) \).

Based on the relationship between \( e_i \) and \( e_j \), similar to Yao and Li \cite{22}, we can rewrite model (1) as

\[
\begin{align*}
Y_{i1} & = X_{i1}^T \beta_0 + e_{i1}, \\
Y_{ij} & = X_{ij}^T \beta_0 + c_{j1} e_{i1} + c_{j2} e_{i2} + \cdots + c_{jj-1} e_{jj-1} + e_{ij}, \ i = 1, \ldots, n, \ j = 2, \ldots, m, \tag{3}
\end{align*}
\]

where \( \text{var}(e_{ij}) = b_{jj}^2, \ j = 1, \ldots, m \).

**Remark 1** If the coefficients \( \{c_{jk}, j = 1, \ldots, m, \ k = 1, \ldots, j - 1\} \) and the variances of \( e_{ij}, \ \{b_{jj}^2, j = 1, \ldots, m\} \) can be consistently estimated, then a consistent estimate of the covariance matrix can be obtained according to Cholesky decomposition.

Note that the errors \( e_{ij} \)'s are unknown, but they can be predicted by \( \hat{e}_{ij} = Y_{ij} - X_{ij}^T \hat{\beta}_f \), where \( \hat{\beta}_f \) is obtained by solving the following robust estimating equation based on the first observation of each subject which are independent.

\[
\sum_{i=1}^{n} X_{i1} W_{i1} \psi \left( \frac{Y_{i1} - X_{i1}^T \hat{\beta}_f}{b_1} \right) = 0, \tag{4}
\]

where \( W_{i1} = w(x_{i1}) \) with \( w(\cdot) \) being a function of covariates which is used to downweight the impact of leverage points, \( \psi(\cdot) \) is a bounded score function which is used to reduce the effects of outliers in the response and is chosen to be the Huber function \( \psi(x) = \min\{c, \max\{-c, x\}\} \). The tuning parameter \( c \) is typically selected to balance estimation efficiency and robustness. We refer to Wang et al. \cite{21} for details on this selection. Here similar to He et al. \cite{5}, we choose \( c = 1.5 \). Moreover, following Sinha \cite{19}, the weighting function \( w(x_{i1}) \) is chosen to be \( \min\{1, b_0/((x_{i1} - m_x)^TS_x^{-1}(x_{i1} - m_x))^{r/2}\} \) with \( r \geq 1 \) which is taken to be 1 in the latter simulation studies and real data analysis, where \( b_0 \) is chosen as the 95th percentile of Chi-square distribution with degrees of freedom equal to the dimension of \( x_{i1} \), and \( m_x \) and \( S_x \) are some robust estimates of location and scale of \( x_{i1} \), such as minimum volume ellipsoid estimates of
Let θ can be shown to be consistent and asymptotic normality. Under some regularity conditions assumed in Section 4, the solution of estimating Equation (7) can be properly estimated and the consistency of the following proposed robust estimator can be achieved.

Remark 2 Since estimating Equation (4) only uses the first observation of each subject, values of each covariate are required to be different at the first observation to ensure that the regression parameters β in Equation (4) are estimable.

Then, replacing ε ij in Equation (3) with ˆε ij, we have

\[ Y_{ij} \approx X_{ij}^T \hat{\beta}_0 + c_{j1} \hat{\epsilon}_{i1} + c_{j2} \hat{\epsilon}_{i2} + \cdots + c_{ji-1} \hat{\epsilon}_{ij-1} + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 2, \ldots, m. \]  

(5)

Let \( \xi_0 = (c_{21}, c_{31}, c_{32}, \ldots, c_{m,m-1})^T \), \( \hat{Z}_{ij} = (0_{(j-2)(j-1)/2}, \hat{\epsilon}_{i1}, \ldots, \hat{\epsilon}_{ij-1}, 0_{(m-1)m/2-(j-1)/2})^T \), \( \hat{Z}_{i}^{(-1)} = (\hat{Z}_{i2}, \ldots, \hat{Z}_{im})^T \) and \( \hat{D}_{i}^{(-1)} = (X_i^{(-1)}, \hat{Z}_{i}^{(-1)}) \) with \( X_i^{(-1)} = (X_{i2}, \ldots, X_{im})^T \). Combining \( y_{i1} = x_{i1}^T \beta_0 + \epsilon_{i1} \) and Equation (5), we can write the model in the matrix format:

\[ Y_i \approx \hat{D}_i \hat{\theta}_0 + \epsilon_i, \quad i = 1, \ldots, n, \]  

(6)

where \( \hat{\theta}_0 = (\hat{\beta}_0^T, \xi_0^T)^T \) and \( \hat{D}_i = (D_{i1}, \hat{D}_{i}^{(-1)})^T \) with \( D_1 = (X_{11}^T, 0_{m(m-1)/2})^T \). Then making use of all the observed data and utilizing the common idea for construction of robust estimating equation similar to Equation (4), the estimator of \( \theta_0 \), ˆθ = ( ˆβ T, ˆξ T) T can be obtained by solving the robust estimating equation:

\[ \sum_{i=1}^{n} \hat{D}_i^2 V_{ij}^{-1/2} W_i R_i \psi(V_{ij}^{-1/2}(Y_i - \hat{D}_i \hat{\theta})) = 0, \]  

(7)

where \( V_i = \text{diag}(b_{i1}^2, \ldots, b_{im}^2) \), \( R_i = \text{diag}(R_{i1}, \ldots, R_{im}) \) and \( W_i = \text{diag}(w_{i1}, \ldots, w_{im}) \) are weighting matrix used to downweight the impact of outliers in the combined covariates \( \hat{D}_i \) with the same weighting function as used in estimating Equation (4). In estimating Equation (7), the function \( \psi \) is applied to each component of the vector \( V_{ij}^{-1/2}(Y_i - \hat{D}_i \hat{\theta}) \).

Remark 3 Under the missing mechanism of MAR and the assumption that \( \epsilon_i \) follow elliptical distribution, it can be shown that \( E(R_i \psi(r_{ij})) = 0 \) with \( r_{ij} = (Y_{ij} - \hat{D}_{ij}^T \hat{\theta}_0)/b_{ij} \). This result is presented in Lemma 2 in the supplementary online material. Then, \( E(R_i \psi(V_{ij}^{-1/2}(Y_i - \hat{D}_i \hat{\theta}_0))) = 0. \) Under some regularity conditions assumed in Section 4, the solution of estimating Equation (7) can be shown to be consistent and asymptotic normality.

With the estimate ˆξ, the variances of random variables \( e_{ij} \) can be robustly estimated using the median absolute deviation (MAD) based on the predicted value of \( e_{ij} \). Then, we can obtain the estimate of the covariance matrix which is denoted by \( \hat{\Sigma}_0 \).

We summarize the estimation procedure in the following.

Step 1. Based on the first observation of each subject, solving the robust estimating equation (4) to obtain a robust estimate of \( \beta_0, \hat{\beta}_f \). And then obtain the predicted value \( \hat{e}_{ij} = Y_{ij} - X_{ij}^T \hat{\beta}_f \) for observed response \( Y_{ij} \).

Step 2. Based on model (6), applying the robust estimating equation (7) to obtain the proposed robust estimator of \( \theta_0 = (\hat{\beta}_f^T, \xi_0^T)^T, \hat{\theta} = (\hat{\beta}_f^T, \hat{\xi}_f^T)^T \). Furthermore, we can obtain the predicted values \( \hat{e}_{ij} = Y_{ij} - \hat{D}_{ij}^T \hat{\theta} \) and obtain the robust estimates of the variance \( b_{ij}^2 \) through the median absolute deviation.
Step 3. With the estimate $\hat{\beta}$ and $\hat{\sigma}^2$ and according to the Cholesky decomposition, the estimate of covariance matrix $\hat{\Sigma}_0$ is obtained.

Through the above procedure, the proposed robust estimator of both mean and covariance can be achieved.

From the above estimation procedure, we summarize some features of the proposed method as follows. First, robust techniques such as the robust estimating equations are used to reduce the impact of outliers in the data. Second, based on the rewritten regression model (3), the proposed method can achieve consistent estimates under MAR without specification of the model for the missing data process. Third, through the Cholesky decomposition, the common regression model (1) can be rewritten as (3) which incorporating the covariance parameters as a part of regression parameters. Then, directly applying the robust estimating equation for the rewritten model (3) without further modeling the covariance parameters, simultaneous estimation of both mean and covariance is realized.

4. Asymptotic properties

In this section, we will show the large-sample properties of the proposed estimator under some regularity conditions. Denote $E_n(\beta) = \sum_{i=1}^n X_i W_i \psi((Y_i - X_i^T \beta)/b_1)$ and $U_n(\theta) = \sum_{i=1}^n D_i^T V_i^{-1/2} W_i R_i \psi(V_i^{-1/2}(Y_i - D_i \theta))$, where $D_i$ denotes $\hat{D}_i$ evaluated at $(\epsilon_1, \ldots, \epsilon_{j-1})^T$. To establish the asymptotic normality of the proposed robust estimator $\hat{\theta}$, some regularity conditions are assumed.

(C.1) The parametric space $\Upsilon$ is compact, for any $\beta \in \Upsilon$. With probability one, $(1/n)E_n(\beta)$ converges uniformly to a non-stochastic function $E(\beta)$ which has a unique zero at $\beta_0 \in \Upsilon$.

(C.2) There is a neighbor $G(\theta_0)$ of $\theta$ on which with probability one, $U_n(\theta)$ is continuously differentiable.

(C.3) With probability one, $(1/n)P_n \rightarrow P$ and $(1/n)Q_n \rightarrow Q$ for some positive definite matrices $P$ and $Q$ where $P_n = \sum_{i=1}^n X_i W_i \psi((Y_i - X_i^T \beta_0)/b_1)X_i^T$, $Q_n = \sum_{i=1}^n A_i A_i^T$ and $A_i = X_i W_i \psi((Y_i - X_i^T \beta_0)/b_1)$.

(C.4) The parametric space $\Theta$ is compact, for any $\theta \in \Theta$. With probability one, $(1/n)U_n(\theta)$ converges uniformly to a non-stochastic function $U(\theta)$ which has a unique zero at $\theta_0 \in \Theta$.

(C.5) With probability one, $(1/n)S_n \rightarrow S$ and $(1/n)K_n \rightarrow K$ for some positive definite matrices $S$ and $K$ where $S_n = \sum_{i=1}^n D_i^T V_i^{-1/2} W_i R_i \psi(r_i)D_i$, $K_n = \sum_{i=1}^n B_i B_i^T$ and $B_i = (\sum_{j=1}^n D_{ij} (w_{ij} R_{ij}/b_j) \psi(r_{ij})) - (\sum_{j=1}^n \sum_{k=1}^m D_{ij} (w_{ij} R_{ij}/b_j) \psi(r_{ij}) \sum_{k=1}^{j-1} S_{ik} X_{ik}^T) \{\sum_{i=1}^n X_i W_i \psi(r_{ij})\}^{-1} X_i W_i \psi(r_{ij})$, $r_i = (r_{i1}, \ldots, r_{im})^T$ and $r_j = (Y_j - D_j^T \theta)/b_j$.

Under Conditions (C.1)–(C.3), it can be shown that $\hat{\beta}_j$ has asymptotic normality which is presented in Lemma 1 in the supplementary online material. Further combining Conditions (C.4) and (C.5), the asymptotic normality of the proposed estimator $\hat{\theta}$ can be achieved. Moreover, Condition (C.2) that $U_n(\theta)$ is continuously differentiable is also assumed in Sinha [19]. As indicated in Sinha [19], the derivative of the Huber function $\psi(\cdot)$ is not continuous at the points $\pm c$, and as a remedy, it is possible to smooth the function $\psi(\cdot)$ to get a continuous derivative. However, the use of such smoothed $\psi(\cdot)$ would not bring dramatic change for the parameter estimation.

Theorem 1 Suppose that Conditions (C.1)–(C.5) hold. Then, we have $n^{1/2}(\hat{\theta} - \theta_0) \rightarrow N(0, \Omega)$, where $\Omega = S^{-1} K S^{-1}$. 


A brief proof is shown in the supplementary online material. Note that to derive the asymptotic covariance matrix \( \Omega_1 \), the variation induced by the estimator \( \hat{\beta}_f \) has to be considered, which is reflected in the matrix \( K_n \) in Condition (C.5).

Following Qin and Zhu [13], \( K \) and \( S \) can be consistently estimated by \( 1/n \) times

\[
\hat{K}_n = \sum_{i=1}^{n} \hat{B}_i \hat{B}_i^T,
\]

with \( \hat{B}_i = (\sum_{j=1}^{m} \hat{D}_{ij} (w_{ij} R_{ij} / \hat{b}_j) \psi(\hat{r}_{ij})) - (\sum_{i=1}^{n} \sum_{j=1}^{m} \hat{D}_{ij} (w_{ij} R_{ij} / \hat{b}_j) \psi(\hat{r}_{ij})) \sum_{k=1}^{j-1} \hat{\zeta}_{jk} X_k^T \{ \sum_{i=1}^{n} X_{ij} w_{ij} \hat{\psi}(\hat{r}_{ij}) X_i^T \}^{-1} X_i w_{ij} \hat{\psi}(\hat{r}_{ij}) \hat{r}_i = (\hat{r}_{i1}, \ldots, \hat{r}_{im})^T \) and \( \hat{r}_{ij} = (Y_{ij} - \hat{D}_{ij} \hat{\theta}) / \hat{b}_j \).

And

\[
\hat{S}_n = \sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1/2} W_i R_i \hat{\psi}(\hat{r}_i) \hat{D}_i,
\]

with \( \hat{V}_i = \text{diag}(\hat{b}_1^2, \ldots, \hat{b}_m^2) \).

Thus, statistical inference can be made on \( \theta_0 \).

5. Simulation studies

In this section, simulation studies are conducted to investigate the performance of the proposed estimator. We compare the proposed method with robust inverse probability weighting GEE (IPW–GEE) method [14] and robust complete-case GEE (CC–GEE) method [5] as well as their non-robust versions. The robust methods reduce to their corresponding non-robust versions when \( \psi(x) = x \) and \( w(x) = 1 \). Furthermore, IPW quadratic inference function (QIF) [15] estimator which incorporates the IPW into the scores for dealing with missing data and the estimator \( \hat{\beta}_f \) which is based on the first observations of the subjects are also calculated for comparison. Under each simulation setting, 500 replications are conducted.

The regression model considered in the simulation study is

\[
Y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \beta_2 X_{2,ij} + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,
\] (8)

where \( X_{1,ij} \) is generated from standard normal distribution and \( X_{2,ij} \) follows binary distribution with success probability 0.6, \( \beta = (\beta_0, \beta_1, \beta_2)^T \) is taken to be \( (1, 1, 1)^T \), \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T \) follows a multivariate normal distribution with mean zero and covariance matrix \( \Sigma \) which is taken to be exchangeable (EX), one-order autoregressive (AR1) and unstructured (UN) structures, respectively. And in the case of EX and AR1, the correlation parameter is taken to be 0.6. In the UN case, the covariance matrix is taken to be

\[
\begin{pmatrix}
1 & 0.7 & 0.7 & 0.4 & 0.2 & 0.2 \\
0.7 & 1 & 0.7 & 0.4 & 0.2 \\
0.7 & 0.7 & 1 & 0.7 & 0.4 \\
0.4 & 0.7 & 0.7 & 1 & 0.7 \\
0.2 & 0.4 & 0.7 & 1 & 0.7 \\
0.2 & 0.2 & 0.4 & 0.7 & 1 \\
\end{pmatrix}.
\]

The number of subjects \( n \) is taken to be 100 and the number of the observations within each subject \( m \) is taken to be 6.
Denote $\lambda_{ij} = P(R_{ij} = 1|R_{ij-1} = 1, \tilde{Y}_{ij})$, where $\tilde{Y}_{ij} = (Y_{i1}, \ldots, Y_{i,j-1})$. A common logistic model is used to model the dropout process

$$\ln \left( \frac{\lambda_{ij}}{1 - \lambda_{ij}} \right) = \gamma_0 + \gamma_1 Y_{i,j-1},$$

where $\gamma = (\gamma_0, \gamma_1)^T$ is taken to be $(1, 1)^T$ and $(0.1, 1)^T$, which will result in about 25% and 40% missing rates, respectively.

We calculate the bias, standard error (SE) and mean square error (MSE) for the estimators of parameters in the mean. To investigate the performance of the proposed method in estimation of the covariance, we calculate the entropy loss

$$\Delta_E(\Sigma, \hat{\Sigma}) = \text{trace}(\Sigma^{-1} \hat{\Sigma}) - \log |\Sigma^{-1} \hat{\Sigma}| - m$$

and quadratic loss

$$\Delta_Q(\Sigma, \hat{\Sigma}) = \text{trace}(\Sigma^{-1} \hat{\Sigma} - I)^2,$$

which can reflect the accuracy in estimation of the covariance matrix where $\Sigma$ is the true covariance matrix and $\hat{\Sigma}$ is its estimate. When $\hat{\Sigma} = \Sigma$, the loss is zero. Smaller loss indicates more estimation accuracy. In our simulation studies, the exchangeable structure is taken to be the working correlation structure for the IPW–GEE and CC–GEE methods.

To study the robustness of the proposed estimator, we consider two ways to generate outliers.

C1. Six completely observed points are randomly chosen. Their covariates $X_{ij}$ are replaced by $X_{ij} - 2$ and their responses $Y_{ij}$ are replaced by $Y_{ij} + 2$.

![Figure 1. Simulation results under the exchangeable correlation structure in the study with 25% missing rate. The numbers 0, 1 and 2 in the x-axis represent the cases of no contamination, Contaminations 1 and 2; R–P, R–IPW and R–CC represent the proposed robust method, robust IPW method and robust complete-case GEE method; NR–P, NR–IPW and NR–CC represent the corresponding non-robust version. IPW–QIF represents the IPW quadratic inference function method and CC–F represents the CC method based on the first observation of each subject.](image-url)
C2. Twelve completely observed points are randomly chosen. The way to create outliers is the same as that in C1.

Figures 1–4 and Tables 1 and 2 summarize the simulation results. The simulation results of the bias and MSE under AR1 correlation structure are similar to those under the EX structure, and thus are omitted for brevity. It can be observed from Figures 1 to 4 that in the case of no outliers, the proposed robust method is able to provide consistent estimates for the mean and perform almost equally well with its non-robust version. Furthermore, it generally shows better performance than other estimators in comparison with smaller MSEs. The IPW–GEE and IPW–QIF methods can provide consistent estimates but with larger SEs induced by the IPW. And it is not surprising that the CC–GEE method produces inconsistent estimates due to missing data. From Table 1, the proposed robust estimator for the covariance seems to provide less accurate estimation. This is a premium which one needs to pay for using the robust method when there is, in fact, no outliers in the data. However, in the case of outliers, it can be found that the gain from the use of robust method is huge. The non-robust estimators are heavily influenced by the outliers and all the robust estimators provide smaller biases than those by the non-robust estimators. Specifically, the proposed robust estimator for the mean outperforms the other estimators with smallest MSE, and the proposed robust estimator for the covariance show its robustness with much more estimation accuracy than its non-robust version. In summary, the proposed robust estimator shows its strength in dealing with both robustness and missingness as well as its advantage over the IPW–GEE, IPW–QIF and CC–GEE estimators.

![Figure 2](image-url)  
Figure 2. Simulation results under the unstructured correlation structure in the study with 25% missing rate. For other notations, see Figure 1.
Figure 3. Simulation results under the exchangeable correlation structure in the study with 40% missing rate. For other notations, see Figure 1.

Figure 4. Simulation results under the unstructured correlation structure in the study with 40% missing rate. For other notations, see Figure 1.
Table 1. Simulation results on the covariance estimation in the study.

|                | EL    | QL    | EL    | QL    |
|----------------|-------|-------|-------|-------|
| **R**          |       |       |       |       |
| 25% missing rate |       |       |       |       |
| NC             | 0.52  | 0.46  | 0.36  | 0.17  |
| AR             | 0.54  | 0.49  | 0.37  | 0.18  |
| UN             | 0.69  | 1.22  | 0.51  | 0.79  |
| C1             | 0.60  | 0.89  | 1.07  | 7.58  |
| AR             | 0.63  | 0.90  | 1.04  | 6.57  |
| UN             | 2.72  | 16.95 | 6.82  | 108.71|
| C2             | 0.75  | 2.16  | 2.10  | 26.39 |
| AR             | 0.83  | 2.27  | 2.02  | 22.64 |
| UN             | 5.30  | 57.41 | 14.05 | 379.69|
| 40% missing rate |       |       |       |       |
| NC             | 0.84  | 0.63  | 0.57  | 0.24  |
| AR             | 0.87  | 0.62  | 0.60  | 0.25  |
| UN             | 0.99  | 1.40  | 0.72  | 0.88  |
| C1             | 0.95  | 1.42  | 1.64  | 12.03 |
| AR             | 1.03  | 1.72  | 1.60  | 10.67 |
| UN             | 3.61  | 25.36 | 8.58  | 156.46|
| C2             | 1.17  | 3.53  | 3.02  | 40.65 |
| AR             | 1.29  | 3.81  | 2.91  | 35.14 |
| UN             | 6.88  | 89.33 | 17.72 | 567.58|

Notes: NC, no contamination; C1, Contamination 1; C2, Contamination 2; EL, entropy loss function; QL, quadratic loss function.

Table 2 presents the average estimated SEs based on 500 replications. It can be found that the empirical SEs are close to the estimated ones which indicates that the large-sample estimate of the variance for the proposed estimator is quite acceptable.

6. Application to a real data

For illustration, the proposed method is applied to a real data set. The data set is a subsample of the database from the National longitudinal Survey conducted by the US Department of Labor [1]. The sample consists of 716 women who were all interviewed for five times at scheduled years and have complete observation values.

It is of interests to explore the factors which significantly affect the wage of the women. The response variable is log transformation of wage. The explanatory variables include Education (EDU), total labor force experience (EXP), tenure in current job (TENU) and dummy variables SOUTH (1, if the woman is from the South; 0, otherwise), BLACK (1, if the woman is black; 0, otherwise) and UNION (1, if the woman is the union member; 0, otherwise).

We fit the data through a linear regression model and first analyze the complete data set using the robust GEE method [5]. The results are summarized in Table 3. Based on the complete data, all the covariate effects estimated by the robust GEE method are significant at level 0.05. Furthermore, EDU, UNION, EXP and TENU show positive effects on the wage, while SOUTH and BLACK present negative effects. But the estimated TENU effect by the non-robust method is insignificant, which indicates the influence induced by some outliers in the data. To further
Table 2. Simulation results on SEs of the proposed estimator in study.

|                | $\beta_0$ SE | $\beta_0$ ESE | $\beta_1$ SE | $\beta_1$ ESE | $\beta_2$ SE | $\beta_2$ ESE |
|----------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 25% missing rate |              |              |              |              |              |              |
| NC             | EX 0.0990    | 0.1010       | 0.0371       | 0.0379       | 0.0772       | 0.0774       |
|                | AR 0.0910    | 0.0922       | 0.0402       | 0.0410       | 0.0844       | 0.0837       |
|                | UN 0.0919    | 0.0950       | 0.0308       | 0.0324       | 0.0644       | 0.0663       |
| C1             | EX 0.0987    | 0.1026       | 0.0386       | 0.0430       | 0.0826       | 0.0816       |
|                | AR 0.0896    | 0.0939       | 0.0422       | 0.0465       | 0.0896       | 0.0882       |
|                | UN 0.0922    | 0.0969       | 0.0351       | 0.0391       | 0.0752       | 0.0744       |
| C2             | EX 0.0988    | 0.1042       | 0.0405       | 0.0484       | 0.0853       | 0.0857       |
|                | AR 0.0906    | 0.0960       | 0.0439       | 0.0523       | 0.0927       | 0.0926       |
|                | UN 0.0936    | 0.0992       | 0.0383       | 0.0455       | 0.0829       | 0.0806       |
| 40% missing rate |              |              |              |              |              |              |
| NC             | EX 0.1038    | 0.1064       | 0.0439       | 0.0434       | 0.0946       | 0.0897       |
|                | AR 0.0983    | 0.0998       | 0.0466       | 0.0466       | 0.0986       | 0.0959       |
|                | UN 0.0968    | 0.1011       | 0.0374       | 0.0376       | 0.0781       | 0.0788       |
| C1             | EX 0.1048    | 0.1085       | 0.0460       | 0.0505       | 0.1012       | 0.0950       |
|                | AR 0.0977    | 0.1027       | 0.0485       | 0.0540       | 0.1056       | 0.1020       |
|                | UN 0.1000    | 0.1038       | 0.0434       | 0.0463       | 0.0933       | 0.0876       |
| C2             | EX 0.1069    | 0.1113       | 0.0491       | 0.0579       | 0.1066       | 0.1012       |
|                | AR 0.1019    | 0.1060       | 0.0510       | 0.0620       | 0.1118       | 0.1077       |
|                | UN 0.1014    | 0.1076       | 0.0466       | 0.0549       | 0.1017       | 0.0959       |

Notes: SE, standard error; ESE, estimated SE using large-sample approximation; EX, exchangeable correlation structure; AR, one-order autoregressive correlation structure; UN, unstructured correlation structure.

To explore possible outliers in the data set, we calculate the weights

$$\sigma_{ij} = \frac{w(x_{ij}) \psi(\tilde{r}_{ij})}{\tilde{r}_{ij}},$$

where $\tilde{r}_{ij} = (Y_{ij} - X_{ij}\hat{\beta})/\tilde{\phi}$ is a consistent estimator of the variance of random errors, $w(\cdot)$ and $\psi$ are defined in Section 3. Small weights indicate that the corresponding observations are possible outliers. We investigate 10 observations with smallest weights and find that these observations seems to seriously deviate from the previous findings on the effects of the covariates. For example, observation 1604 has extremely low EDU value 6 which is below 0.5% quantile of the EDU values in the sample. Its EXP and TENU values are close to their medians in the sample, and its corresponding subject is black and from the South. According to the previously analysis, the observation with such covariate values is expected to have low wage. But observation 1604 has extremely high wage value 4.0 which is beyond 99.5% quantile of the wage values in the sample.

To illustrate the application of the proposed method for missing data, we artificially produce dropouts for the subjects based on the following dropout model:

$$\ln \left( \frac{\lambda_{ij}}{1 - \lambda_{ij}} \right) = \gamma_0 + \gamma_1 Y_{ij-1} + \gamma_2 \text{TENU}_{ij},$$
Table 3. Regression coefficient estimates in the analysis of the real data.

|        | Based on incomplete data | Based on complete data |
|--------|--------------------------|------------------------|
|        | R–P | R–IPW | R–CC | NR–P | NR–IPW | NR–CC | IPW–QIF | R–GEE | NR–GEE |
| EDU    | 0.0668 (0.0050)  | 0.0737 (0.0057) | 0.0705 (0.0050) | 0.0648 (0.0054) | 0.0732 (0.0057) | 0.0698 (0.0055) | 0.0756 (0.0058) | 0.0751 (0.0051) | 0.0739 (0.0054) |
| SOUTH  | −0.1230 (0.0266) | −0.0495 (0.0414) | −0.0918 (0.0274) | −0.0923 (0.0277) | 0.0667 (0.0840) | −0.0441 (0.0355) | −0.0929 (0.0293) | −0.1161 (0.0249) | −0.0677 (0.0317) |
| BLACK  | −0.0768 (0.0282) | −0.1092 (0.0316) | −0.1009 (0.0295) | −0.0894 (0.0293) | −0.1469 (0.0528) | −0.1184 (0.0321) | −0.1022 (0.0291) | −0.1108 (0.0282) | −0.1193 (0.0301) |
| UNION  | 0.0877 (0.0235)  | 0.0834 (0.0206)  | 0.0817 (0.0176)  | 0.1054 (0.0243)  | 0.0786 (0.0219)  | 0.0773 (0.0179)  | 0.0882 (0.0196)  | 0.0803 (0.0152)  | 0.0762 (0.0161)  |
| EXP    | 0.0258 (0.0047)  | 0.0234 (0.0054)  | 0.0223 (0.0051)  | 0.0262 (0.0047)  | 0.0242 (0.0069)  | 0.0244 (0.0070)  | 0.0253 (0.0041)  | 0.0268 (0.0022)  | 0.0279 (0.0024)  |
| TENU   | 0.0095 (0.0037)  | 0.0076 (0.0045)  | 0.0081 (0.0046)  | 0.0100 (0.0038)  | 0.0050 (0.0060)  | 0.0051 (0.0065)  | 0.0058 (0.0034)  | 0.0058 (0.0020)  | 0.0028 (0.0021)  |

Notes: The figures within the parenthesis are SEs.
Table 4. Correlation matrix estimates in the analysis of the real data.

|          | R–P        | NR–P        |
|----------|------------|------------|
|          | 1.0000     | 0.7865     |
| 1.0000   | 0.9299     | 0.6916     |
| 0.9299   | 0.8690     | 0.7982     |
| 0.8690   | 0.9165     | 0.7982     |
| 0.8231   | 0.8612     | 1.0000     |
| 0.8309   | 0.8965     | 0.8159     |
| 0.8309   | 0.9004     | 1.0000     |

Notes: R–P, robust proposed method; NR–P, non-robust version.

where $\gamma = (\gamma_0, \gamma_1, \gamma_2)^T$ is taken to be $(1, -0.8, 0.8)^T$, which will result in about 20% missing rate. We analyze the data with dropouts using the proposed method, inverse probability weighting GEE (IPW–GEE), complete-case GEE (CC–GEE) and IPW–QIF methods, respectively.

Table 3 also presents the results of the estimates for the regression coefficient based on the incomplete data. The proposed method provides consistent significant results with the robust GEE method based on the complete data. Moreover, there exists obvious numerical difference between the proposed robust method and its non-robust version although they both get consistent significant results. However, the other methods miss at least one significant covariate. Like the findings in the simulation studies, the IPW method is sometimes unstable and may result in larger standard errors due to the weighting and the CC method which ignore the dropouts generally produce inconsistent estimates.

Table 4 shows the estimate for the correlation matrix based on the estimated covariance matrix. The estimated correlation matrix by the proposed robust method indicates higher correlation among the five observations within each women.

In summary, the proposed method is helpful to provide more reliable and accurate results due to its strength in dealing with outliers and missing data, and enables us to gain more useful information inherited in the data for its ability to achieve estimation of both mean and covariance.

7. Discussions and future work

In this paper, we develop a robust approach for estimation of both mean and covariance components for longitudinal data with dropouts. The proposed method is against outliers in the data, and is able to achieve consistent estimates of the parameters of interest in the case of dropouts without need to specify the model for the missing data process. To implement the proposed method, the values of each covariate at the first observation are required to be different. To handle this issue, we can use the estimate proposed in Qu et al. [16] to get an initial estimate of $\hat{\beta}_0$ instead of $\hat{\beta}_f$. In this way, the final estimate is still consistent and is expected to possess some robustness. However, the method proposed by Qu et al. [16] is not robust which may influence the prediction of $\epsilon_j$. Therefore, the performance of this alternative method needs to be assessed and can be investigated in the future study.

We here assume that all the subjects share the same covariance matrix which can be generalized to more general cases, where the covariance matrix depends on some covariates using
the same idea through modeling the relationship between the covariance parameters and some covariates. Further work and additional efforts are necessary to study this topic.

The breakdown properties of the proposed robust estimator is yet another interesting topic. Zhang and Li [25] studied the breakdown properties of location M-estimators for independent data. However, it may be more complicated here since we study the robust estimation of the mean and covariance simultaneously for longitudinal data with missing values. This requires further careful investigation as a future study.

Acknowledgements

The authors are grateful to two referees for their constructive suggestions that largely improve the presentation of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially supported by the National Nature and Science Foundation of China [11371100, 11271080], the Scientific Research Foundation for the Returned Overseas Chinese Scholars and Shanghai Leading Academic Discipline Project, Project number: B118.

Supplemental data and research materials

Supplemental data for this article can be accessed at 10.1080/02664763.2014.999033.

References

[1] R. Carter Hill, W.E. Griffiths, and G.C. Lim, *Principles of Econometrics*, 3rd ed., Wiley, Hoboken, NJ, 2008.
[2] B. Chen and X.-H. Zhou, *Generalized partially linear models for incomplete longitudinal data in the presence of population-level information*, Biometrics 69 (2013), pp. 386–395.
[3] T.P. Garcia, P. Kohli, and M. Pourahmadi, *Regressograms and mean–covariance models for incomplete longitudinal data*, Amer. Statist. 66 (2012), pp. 85–91.
[4] E. Gómez, M.A. Gómez-Villegas, and J.M. Marín, *A survey on continuous elliptical vector distributions*, Rev. Mat. Complut. 16 (2003), pp. 345–361.
[5] X. He, W.K. Fung, and Z. Zhu, *Robust estimation in generalized partial linear models for clustered data*, J. Amer. Statist. Assoc. 100 (2005), pp. 1176–1184.
[6] C. Leng, W. Zhang, and J. Pan, *Semiparametric mean–covariance regression analysis for longitudinal data*, J. Amer. Statist. Assoc. 105 (2010), pp. 181–193.
[7] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., John Wiley, New York, 2002.
[8] J. Mao, Z. Zhu and W.K. Fung, *Joint estimation of mean–covariance model for longitudinal data with basis function approximations*, Comput. Statist. Data Anal. 55 (2011), pp. 983–992.
[9] P.J. Meyer, *Applied Measurement with JMetrik*, 1st ed., Routledge, New York, 2014.
[10] M. Pourahmadi, *Joint mean–covariance models with applications to longitudinal data: Unconstrained parameterisation*, Biometrika 86 (1999), pp. 677–690.
[11] M. Pourahmadi, *Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix*, Biometrika 87 (2000), pp. 425–435.
[12] J.S. Preisser, K.K. Lohman, and P.J. Rathouz, *Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random*, Stat. Med. 21 (2002), pp. 3035–3054.
[13] G. Qin and Z. Zhu, *Robust estimation in generalized semiparametric mixed models for longitudinal data*, J. Multivariate Anal. 98 (2007), pp. 1658–1683.
[14] G.Y. Qin, Z.Y. Zhu, and W.K. Fung, *Robust Estimation of Generalized Partially Linear Model for Longitudinal Data with Dropouts*. Submitted to Ann. Inst. Stat. Math., under revision, 2014.
[15] A. Qu, B.G. Lindsay, and B. Li, *Improving generalised estimating equations using quadratic inference functions*, Biometrika 87 (2000), pp. 823–836.

[16] A. Qu, B.G. Lindsay, and L. Lu, *Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random*, J. Amer. Statist. Assoc. 105 (2010), pp. 194–204.

[17] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Analysis of semiparametric regression models for repeated outcomes in the presence of missing data*, J. Amer. Statist. Assoc. 90 (1995), pp. 106–121.

[18] P.J. Rousseeuw and B.C. van Zomeren, *Unmasking multivariate outliers and leverage points*, J. Amer. Statist. Assoc. 85 (1990), pp. 633–639.

[19] S.K. Sinha, *Robust analysis of generalized linear mixed models*, J. Amer. Statist. Assoc. 99 (2004), pp. 451–460.

[20] N. Wang, *Marginal nonparametric kernel regression accounting within-subject correlation*, Biometrika 90 (2003), pp. 29–42.

[21] Y.-G. Wang, X. Lin, M. Zhu, and Z. Bai, *Robust estimation using the Huber function with a data-dependent tuning constant*, J. Comput. Graph. Statist. 16 (2007), pp. 468–481.

[22] W Yao and R. Li, *New local estimation procedure for a non-parametric regression function for longitudinal data*, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (2013), pp. 123–138.

[23] H. Ye and J. Pan, *Modelling of covariance structures in generalised estimating equations for longitudinal data*, Biometrika 93 (2006), pp. 927–941.

[24] G.Y. Yi and W. He, *Median regression models for longitudinal data with dropouts*, Biometrics 65 (2009), pp. 618–625.

[25] J. Zhang and G. Li, *Breakdown properties of location M-estimators*, Ann. Statist. 26 (1998), pp. 1170–1189.

[26] X. Zheng, W.K. Fung, and Z. Zhu, *Robust estimation in joint mean–covariance regression model for longitudinal data*, Ann. Inst. Stat. Math. 65 (2013), pp. 617–638.

[27] Z. Zhu, W.K. Fung, and X. He, *On the asymptotics of marginal regression splines with longitudinal data*, Biometrika 95 (2008), pp. 907–917.