应用概率统计　第 34 卷
第 6 期　2018 年 12 月

Chinese Journal of Applied Probability and Statistics
Dec., 2018, Vol. 34, No. 6, pp. 598-612
doi: 10.3969/j.issn.1001-4268.2018.06.005

Efficient Robust Estimation of Mean and Covariance for Longitudinal Data *

FAN Yali⋆　　XU Xiaolin

(*College of Science, University of Shanghai for Science and Technology, Shanghai, 200093, China*)

Abstract: In this article, we develop efficient robust method for estimation of mean and covariance simultaneously for longitudinal data in regression model. Based on Cholesky decomposition for the covariance matrix and rewriting the regression model, we propose a weighted least square estimator, in which the weights are estimated under generalized empirical likelihood framework. The proposed estimator obtains high efficiency from the close connection to empirical likelihood method, and achieves robustness by bounding the weighted sum of squared residuals. Simulation study shows that, compared to existing robust estimation methods for longitudinal data, the proposed estimator has relatively high efficiency and comparable robustness. In the end, the proposed method is used to analyse a real data set.

Keywords: efficient estimation; generalized empirical likelihood; robust

2010 Mathematics Subject Classification: 91G70

Citation: FAN Y L, XU X L. Efficient robust estimation of mean and covariance for longitudinal data [J]. Chinese J Appl Probab Statist, 2018, 34(6): 598–612.

§1. Introduction

Longitudinal data arise when a response variable is measured repeatedly through time for independent subjects. A major character of longitudinal data is the within-subject correlation and between-subject heterogeneity. Many authors consider the estimation of mean parameter and covariance matrix for longitudinal data. Liang and Zeger [1] introduced the technique of generalized estimating equations (GEE) for marginal mean model. Qu et al. [2] showed how to exploit the within-subject correlation based on QIF methods to improve the efficiency of the GEE based estimator. Ye and Pan [3] proposed an approach for joint modeling of mean and covariance structures within the GEE framework. An incomplete list of related works includes [4–8].

*The project was supported by the National Natural Science Foundation of China (Grant No. 11401383).

⋆Corresponding author, E-mail: yalifan@usst.edu.cn.

Received May 15, 2018. Revised September 14, 2018.

On the other hand, robustness study is also very important for longitudinal data analysis, since one outlier in the subject level may generate a set of outliers in the sample due to repeated measurements. Some literatures, e.g., [9–12], developed robust regression methods for estimation on mean and covariance matrix. Qin and Zhu [13] proposed a approach to achieve simultaneously robust estimation of both mean and covariance for longitudinal data with dropouts. However, robustness gains usually entails some loss of efficiency. The main purpose of this article is to improve the efficiency of estimator while ensuring their resistance to outliers in finite samples for longitudinal data analysis.

The empirical likelihood (EL) method, first proposed by Owen [14], is a nonparametric-likelihood-based approach and has attracted a great deal of interests, many authors extended the EL method to longitudinal data analysis, see [15–17]. The EL method enables us to fully employ the information and incorporate side information through constraints for making asymptotically efficient inference. Qin and Lawless [18] linked EL method and estimating equations under moment restriction. They showed that EL estimator is asymptotically efficient if the moment specifications are correct and the likelihood score functions are included as a subset of the restrictions. Bondell and Stefanski [19] proposed a robust estimator in linear regression which has relatively high efficiency compared to other robust estimators by using generalized EL methods.

Motivated by the efficiency of EL methods and the work of [19], we proposed a more efficient estimation method to achieve simultaneously robust estimation of both mean and covariance for longitudinal data. The advantages of the proposed method can be summarized in four points. First, the proposed estimator is robust against possible outliers in the dataset. Second, compared to the existing robust estimators, the proposed estimator has relatively high efficiency in finite sample and comparable outlier resistance. Third, different with the common GEE approach for longitudinal analysis, our method does not need to specify the working correlation structure or model the parameters of the covariance, and thus, the issue of correlation misspecification can be avoid. Fourth, unlike other robust estimating equations based methods, and instead of choosing the tuning threshold constants empirically by hand, our method depend on data driven weights which can be estimated directly under the generalized EL framework.

The rest of this article is organized as follows. In Section 2, we develop the proposed estimator. The algorithm that is used to compute the proposed estimator is discussed in Section 3. Simulation studies are conducted to compare the performance of the proposed estimator with existing ones in Section 4. A real data set analysis is presented in Section 5.

§2. Proposed Method

2.1 Models

Suppose that we have a sample of n subjects with m observations over time for each subject. We will consider the following linear regression model

$$y_{ij} = x_{ij}^{\mathsf{T}}\beta + \epsilon_{ij}, \qquad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m, \tag{1}$$

where y_{ij} is the jth observation on the ith subject, x_{ij} is a p-vector of covariance values, β is a p-dimensional vector of unknown regression coefficients, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{im})^{\mathsf{T}}$ are independently distributed with mean $0_{m\times 1}$ and covariance matrix Σ, which is unknown positive definite matrix.

To simultaneously estimate the parameters in β and Σ, we perform the modified Cholesky decomposition of Σ. According to the positive definite property of Σ, there exists a unique lower triangular matrix C with 1's being the diagonal entries and a unique diagonal matrix D with positive diagonals such that

$$\mathsf{Cov}\,(C\epsilon_i) = C\Sigma C^{\mathsf{T}} = D. \tag{2}$$

Denote $e_i = (e_{i1}, e_{i2}, \ldots, e_{im})^{\mathsf{T}} = C\epsilon_i$ and $D = \mathrm{diag}(d_1^2, d_2^2, \ldots, d_m^2)$, then we have

$$\begin{cases} \epsilon_{i1} = e_{i1}, \\ \epsilon_{ij} = c_{j1}\epsilon_{i1} + c_{j2}\epsilon_{i2} + \cdots + c_{j,j-1}\epsilon_{i,j-1} + e_{ij}, \quad i = 1, 2, \ldots, n,\ j = 2, 3, \ldots, m, \end{cases} \tag{3}$$

where c_{jk} is the negative of the (j,k)-component of C.

Based on relationship (3), and similar to [13], we can rewrite model (1) as

$$\begin{cases} y_{i1} = x_{i1}^{\mathsf{T}}\beta + e_{i1}, \\ y_{ij} = x_{ij}^{\mathsf{T}}\beta + c_{j1}\epsilon_{i1} + c_{j2}\epsilon_{i2} + \cdots + c_{j,j-1}\epsilon_{i,j-1} + e_{ij}, \quad i = 1, 2, \ldots, n,\ j = 2, 3, \ldots, m. \end{cases} \tag{4}$$

Note that in (4), the errors ϵ_{ij}s are unknown, similar to [13], we use the predicted value $\widehat{\epsilon}_{ij} = y_{ij} - x_{ij}^{\mathsf{T}}\widehat{\beta}_j$, where $\widehat{\beta}_j$ is the robust estimator of β which based on the jth observation of the n independent subjects, and $\widehat{\beta}_j$ can be obtained by solving the following robust estimating equations

$$\sum_{i=1}^{n} x_{ij}\omega_{ij}\psi(y_{ij} - x_{ij}^{\mathsf{T}}\beta_j)/d_j = 0, \tag{5}$$

where $\psi(\cdot)$ is chosen to limit the influence on outliers in response, and a common choice is Huber's score function $\psi_c(x) = \min\{c, \max\{-c, x\}\}$ for some constant c, normally chosen to be between 1 and 2. Following [20, 21], we choose the weights $\omega_{ij} = \omega(x_{ij})$ as the Mahalanobis distance in the form

$$\omega_{ij} = \omega(x_{ij}) = \min\Big\{1, \Big[\frac{b_0}{(x_{ij} - m_x)^{\mathsf{T}} S_x^{-1}(x_{ij} - m_x)}\Big]^{r/2}\Big\} \tag{6}$$

with $r \geqslant 1$, and m_x, S_x are some robust estimates of location and scale of x_{ij}. We refer to [10, 20] for details.

Then, we replace ϵ_{ij} in (4) with $\widehat{\epsilon}_{ij}$, and let $c = (c_{21}, c_{31}, c_{32}, \ldots, c_{m,m-1})^{\mathsf{T}}$, $\theta = (\beta^{\mathsf{T}}, c^{\mathsf{T}})^{\mathsf{T}}$, $z_{i1} = (x_{i1}, 0_{m\times(m-1)/2})^{\mathsf{T}}$, and for $j = 2, 3, \ldots, m$, $z_{ij} = (x_{ij}^{\mathsf{T}}, 0_{(j-2)\times(j-1)/2}, \widehat{\epsilon}_{i1}, \widehat{\epsilon}_{i2}, \ldots, \widehat{\epsilon}_{i,j-1}, 0_{(m-1)\times m/2-(j-1)\times j/2})^{\mathsf{T}}$. We can reparameterized model (4) as

$$y_{ij} = z_{ij}^{\mathsf{T}}\theta + e_{ij}, \qquad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m. \tag{7}$$

So, if the coefficients θ and the variance of e_{ij} can be consistently estimated, a consistent estimate of β and Σ can be obtained simultaneously.

2.2 Proposed Estimator

By assigning a probability mass p_{ij} to each observation y_{ij}, we define the element-wise generalized empirical likelihood function as

$$\begin{aligned} L(P, \theta) = \inf_{p_{ij},\theta}\Big\{ &\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\ln(Np_{ij}) \,|\, p_{ij} \geqslant 0,\ \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1, \\ &\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(y_{ij} - z_{ij}^{\mathsf{T}}\theta)z_{ij} = 0,\ \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(y_{ij} - z_{ij}^{\mathsf{T}}\theta)^2 \leqslant \widehat{\sigma}^2_{\mathrm{OLS}}\Big\}, \end{aligned} \tag{8}$$

where $N = n\times m$, $P = (p_{11}, p_{12}, \ldots, p_{1m}, p_{21}, p_{22}, \ldots, p_{2m}, \ldots, p_{n1}, p_{n2}, \ldots, p_{n,m})^{\mathsf{T}}$, $\widehat{\sigma}^2_{\mathrm{OLS}} = N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - z_{ij}^{\mathsf{T}}\widehat{\theta}_{\mathrm{OLS}})^2$, and $\widehat{\theta}_{\mathrm{OLS}}$ is the ordinary least square estimator of model (7).

The estimator reduced from function in (8) is a generalization of the empirical likelihood estimator, which is based on minimization of Cressie-Read discrepancy statistic (see [22–24]). The Cressie-Read minimum discrepancy estimators are based on minimizing the difference between the empirical distribution, that is, the N-dimensional vector with all elements equal to $1/N$, and the estimated weights subject to the restrictions being satisfied.

The proposed estimate method is a two stage estimator. Firstly, by minimizing $L(P, \theta)$ in (8), we can get $\widehat{P} = (\widehat{p}_{11}, \widehat{p}_{12}, \ldots, \widehat{p}_{n,m})$ and $\widetilde{\theta}$. Secondly, the proposed estimator, which is denoted $\widehat{\theta}$, is weighted least squares estimator based on $\widehat{P}$. Specifically,

$$\widehat{\theta} = (Z^{\mathsf{T}}WZ)^{-1}Z^{\mathsf{T}}WY, \tag{9}$$

where $Z = (z_{11}, z_{12}, \ldots, z_{n,m})$, $Y = (y_{11}, y_{12}, \ldots, y_{n,m})^{\mathsf{T}}$, $W = \mathrm{diag}\{\widehat{p}_{11}, \widehat{p}_{12}, \ldots, \widehat{p}_{n,m}\}$.

The advantages of the proposed estimator can be explained from three aspects. First, the weights used in (9) are determined as close as to equal weights as possible, since the weights are obtained by minimizing the distance between equal weights and estimated weights $\widehat{p}_{ij}$. So, we expect the proposed estimator obtain high efficiency when the errors are normally distributed. Second, the proposed estimator obtain robustness by down-weighting observations that do not fit model (7) well. We can see this from the last moment restriction in (8). In fact,

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{p}_{ij}(y_{ij} - z_{ij}^{\mathsf{T}}\widetilde{\theta})^2 \leqslant \widehat{\sigma}_{\mathrm{OLS}}^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - z_{ij}^{\mathsf{T}}\widehat{\theta}_{\mathrm{OLS}})^2 \leqslant \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - z_{ij}^{\mathsf{T}}\widetilde{\theta})^2,$$

since $N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{p}_{ij} = N^{-1}$, it follows that $\sum_{i=1}^{n}\sum_{j=1}^{m}(\widehat{p}_{ij} - N^{-1})(y_{ij} - z_{ij}^{\mathsf{T}}\widetilde{\theta})^2 \leqslant 0$. Thus, the estimated weights $\widehat{p}_{ij}$ and the squared residuals $(y_{ij} - z_{ij}^{\mathsf{T}}\widetilde{\theta})^2$ are negative correlated, and the proposed estimator tends to assign small weights to large squared residuals and vice versa. Third, many robust estimators (see [25, 26]) are also weighted least squares approaches while they construct weights based on an initial measure of outlyingness, however, the proposed estimator use a data-driven weights, since the weights are estimated directly within a generalized empirical likelihood framework.

The proposed estimate method is a generalization of the work of [19], which focused on cross-sectional data analysis. The work of [19] showed that, such two-stage estimator can simultaneously obtain full efficiency under the normal distribution and retaining the asymptotic breakdown point of 1/2. Based on Cholesky decomposition and reparameterization, we generalize this estimate method to longitudinal data analysis under regression model. When compared with existing robust estimator for longitudinal data, the proposed estimator is shown via simulation, to remain higher efficiency and comparable outlier resistance.

Now, we summarize the estimation procedures as following.

Step 1. Based on the jth observation of each subject solving the robust estimating equation (5) to get a robust estimate of β, $\widehat{\beta}_j$, and then get the predicted value $\widehat{\epsilon}_{ij} = y_{ij} - x_{ij}^{\mathsf{T}}\widehat{\beta}_j$.

Step 2. Based on model (7), minimizing the generalized empirical likelihood function (8), to obtain the estimated weights $\widehat{p}_{ij}$, then obtain the proposed estimator of $\theta = (\beta^{\mathsf{T}}, c^{\mathsf{T}})^{\mathsf{T}}$, $\widehat{\theta} = (\widehat{\beta}^{\mathsf{T}}, \widehat{c}^{\mathsf{T}})^{\mathsf{T}}$. Furthermore, we can obtain the predicted values $\widehat{e}_{ij} = y_{ij} - z_{ij}^{\mathsf{T}}\widehat{\theta}$ and get the robust estimates of variance d_j^2 through the median absolute deviation.

Step 3. Based on Cholesky decomposition, and with the estimate $\widehat{c}$ and $\widehat{d}_j^2$, we can get the estimate of covariance matrix $\widehat{\Sigma}$.

§3. Algorithm

To handle the constrained minimization problem in (8), we can set it in Lagrangian form, the Lagrange function is

$$\begin{aligned} L_0(P,\theta,\Lambda) = & \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\ln(Np_{ij}) - \lambda_1\Big(\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} - 1\Big) \\ & - \lambda_2^{\mathsf{T}}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(y_{ij} - z_{ij}^{\mathsf{T}}\theta)z_{ij} - \lambda_3\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}[(y_{ij} - z_{ij}^{\mathsf{T}}\theta)^2 - \sigma_T^2], \end{aligned} \tag{10}$$

where $\Lambda = (\lambda_1, \lambda_2^{\mathsf{T}}, \lambda_3)^{\mathsf{T}}$ is the Lagrange multipliers, and σ_T^2 is a target residual scale determined via a initial estimate which satisfy $\sigma_T^2 \leqslant \widehat{\sigma}_{\mathrm{OLS}}^2$. Taking the derivatives of equation (10) with respect to each p_{ij}, θ, λ_1, λ_2 and λ_3, and setting them to zero reveals $\lambda_2 = 0$ and yields following equations

$$p_{ij} = p_{ij}^*\Big/\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^*, \quad \text{and} \quad p_{ij}^* = \exp\{\lambda_3(y_{ij} - z_{ij}^{\mathsf{T}}\theta)^2 - \sigma_T^2\}, \tag{11}$$

$$\begin{cases} \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} p_{ij}^*[(y_{ij} - z_{ij}^{\mathsf{T}}\theta)^2 - \sigma_T^2] = 0, \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} p_{ij}^*(y_{ij} - z_{ij}^{\mathsf{T}}\theta)z_{ij} = 0. \end{cases} \tag{12}$$

Since we can substitute p_{ij} given in (11), the constrained minimizing problem is reduced to solve equations (12), and find θ and Lagrange multiplier λ_3. However, these equations are not easy to handle since there is no analytical solution. In literature, this problem is generally solved by Newton-type numerical algorithms, see [14,27]. In this paper, similar to [19], we solve alternative saddle-point equations, which are showed equivalent to equations (12).

Denote $J(\lambda_3, \theta) = N^{-1}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\exp\{\lambda_3[(y_{ij} - z_{ij}^{\mathsf{T}}\theta)^2 - \sigma_T^2]\}$ and $J_1(\lambda_3, \theta) = \partial J(\lambda_3, \theta)/\partial\lambda_3$, $J_2(\lambda_3, \theta) = \partial J(\lambda_3, \theta)/\partial\theta$. Based on examining the first and second derivatives of $J(\lambda_3, \theta)$, it can showed that, $J(\lambda_3, \theta)$ is convex in λ_3 for fixed θ and concave in θ for fixed λ_3. We refer to [19] for detail analysis. Thus, we can obtain the estimate of λ_3 and θ by alternating minimization-maximization algorithm. Specifically, we solve the following equations, iterating between θ and λ_3

$$J_1(\lambda_3, \theta) = 0, \quad \text{and} \quad J_2(\lambda_3, \theta) = 0. \tag{13}$$

Expanding (13) and comparing it to (12) reveals their equivalence.

Now, we summarize the computational algorithm in the following.

Step 1. Define $\sigma_T^2 = \min\{\widehat{\sigma}_{\rm LTS}^2, \alpha\widehat{\sigma}_{\rm OLS}^2\}$, where $\widehat{\sigma}_{\rm LTS}^2 = N^{-1}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}(y_{ij} - z_{ij}^{\rm T}\widehat{\theta}_{\rm LTS})^2$, and $\widehat{\theta}_{\rm LTS}$ denote the least trimmed square estimator of model (8), and $0 < \alpha < 1$ is fixed constant near to 1 to exclude the boundary cases. We use $\alpha = 0.95$ in the simulation study.

Step 2. Assign initial values for θ and set a tolerance value ϵ^*.

Step 3. Minimize $J(\lambda_3, \theta)$ with respect to λ_3 for fixed θ by solving $J_1(\lambda_3, \theta) = 0$, and denote the solution by $\lambda_3^*(\theta)$.

Step 4. Maximize profiled function $J(\lambda_3^*(\theta), \theta)$ with respect to θ, and denote the updated estimate by θ^*.

Step 5. Repeat step 3 and step 4 until $\|\theta^* - \theta\| < \epsilon^*$, and denote the resulted estimate of λ_3, θ by $\widetilde{\lambda}_3$, $\widetilde{\theta}$ respectively.

Step 6. Compute the estimate weights by (11).

Step 7. Compute the proposed weighted least square estimate $\widehat{\theta}$ by (9).

Step 8. With the estimate $\widehat{\theta}$ and based on Cholesky decomposition, we obtain the estimate $\widehat{\beta}$ and $\widehat{\Sigma}$.

§4. Simulation Study

In this section, we present a simulation study to investigate the finite sample efficiency and robustness of the proposed estimator. Comparisons are made with two other robust estimators. One is the estimator proposed by Qin and Zhu[13], denoted by Qin-R, which is based on Cholesky decomposition for covariance matrix and robust estimating equations. Although [13] focused on the estimation for longitudinal data with dropouts, the estimation procedures are also adaptive to complete data set, which can be adjusted by simply changing all the missingness indicator to 1. We compute the estimator in complete data case and compare it to the proposed estimator. The other one is the estimator proposed by He et al.[10], denoted by He-R, which is obtained under the robust GEE framework. We also included the non-robust version of these two estimator for comparison, which are denoted by Qin-NR and He-NR respectively.

4.1 Efficiency

To assess efficiency, we generated data from following regression model

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \epsilon_{ij}, \qquad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m, \tag{14}$$

where x_{ij1} and x_{ij2} are generated from standard normal distribution, $\beta = (\beta_0, \beta_1, \beta_2)^\top = (1, 1, 1)^\top$, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{im})^\top$ are generated from multivariate normal distribution with mean zero and covariance matrix Σ, which is set to be independent (Inde), exchangeable (Exch) and one order autoregressive (AR(1)) respectively. The correlation parameter in the case of Exch and AR(1) is taken to be 0.5. In our simulations, the constant r in the weight function (6) and the constant c in the Huber's function $\psi_c(\cdot)$ is chosen to be 1.5, and working correlation structure is taken to be Exch for He-R and He-NR estimators. The number of subject, n, is taken to be 30, 50, 100, 150, and the number of observation for each subject, m, is set to be 4.

Based on 200 replications, we compute the relative efficiency (RE) by the ratio of total mean squared errors (MSE) defined as

$$\mathrm{RE} = \sum_{k=1}^{200} \|\mathrm{MSE}(\widehat{\beta}_k^{\mathrm{MLE}})\|^2 \Big/ \sum_{k=1}^{200} \|\mathrm{MSE}(\widehat{\beta}_k)\|^2, \tag{15}$$

where $\widehat{\beta}_k^{\mathrm{MLE}}$ is the distribution specific maximum likelihood estimator, and $\widehat{\beta}_k$ is the study estimator. We also compute the mean square error for the estimate of β_0, β_1 and β_2 based on the 200 replications. Similar to [13], we assess the performance in estimating the covariance matrix by investigating the entropy loss (EL) and quadratic loss (QL) which are defined as

$$\mathrm{EL} = \mathrm{trace}(\Sigma^{-1}\widehat{\Sigma}) - \ln|\Sigma^{-1}\widehat{\Sigma}| - m; \qquad \mathrm{QL} = [\mathrm{trace}(\Sigma^{-1}\widehat{\Sigma} - I)]^2. \tag{16}$$

It is easy to see the loss is zero when $\widehat{\Sigma} = \Sigma$, so smaller loss indicate more accurate estimation. Since [10] focused on the robust estimator for the mean, we only compare the proposed estimator for covariance matrix to Qin-R and Qin-NR.

Table 1 – Table 3 summarize the simulation results for efficiency comparison. These results are in line with our expectations. From Table 1, it is fair to say the proposed estimator (denoted as P) is highly efficient even with sample size $n = 30$, $n = 50$, since the proposed estimator always enjoy the largest relative efficiency and smallest mean square error. Table 2 shows that, in no outlier case, the proposed estimator for the covariance matrix also show its efficiency with less loss in terms of EL and QL. Table 3 shows the higher efficiency of the proposed estimator by MSE. According to our experience obtained from simulation study, in all cases, the bias are negligible relative to the standard deviation, and hence the mean square error captures the variance comparisons very well. When we compare He-R and Qin-R with their non-robust version estimators, it can be seen that He-R and Qin-R seem to give slightly higher mean square error. This is the price that the robust method must pay when there is, in fact, no outlier in the data, we shall see that under contamination, He-NR and Qin-NR are less robust.

Table 1 Relative efficiency results with respect to β under normal distribution (no contamination cases)

		He-R	He-NR	Qin-R	Qin-NR	P
Inde.	$n = 30$	77	80	66	88	99
	$n = 50$	80	83	70	91	100
	$n = 100$	84	81	73	89	100
	$n = 150$	86	90	82	97	100
Exch.	$n = 30$	84	89	76	90	98
	$n = 50$	87	91	80	96	99
	$n = 100$	87	96	81	92	99
	$n = 150$	90	99	92	98	100
AR(1)	$n = 30$	80	93	82	91	99
	$n = 50$	85	97	87	98	99
	$n = 100$	82	89	88	95	99
	$n = 150$	91	94	92	99	100

Tabel 2 Simulation results for Σ under normal distribution (no contamination cases)

		Inde.		Exch.		AR(1)	
Methods		EL	QL	EL	QL	EL	QL
$n = 30$	Qin-R	0.54	0.36	0.46	0.36	0.64	0.33
	Qin-NR	0.21	0.20	0.29	0.23	0.25	0.28
	P	0.18	0.19	0.28	0.21	0.24	0.25
$n = 50$	Qin-R	0.36	0.28	0.39	0.28	0.39	0.22
	Qin-NR	0.18	0.19	0.18	0.19	0.21	0.19
	P	0.18	0.18	0.14	0.13	0.20	0.16
$n = 100$	Qin-R	0.35	0.27	0.36	0.24	0.36	0.23
	Qin-NR	0.19	0.17	0.15	0.15	0.25	0.16
	P	0.18	0.14	0.18	0.13	0.20	0.15
$n = 150$	Qin-R	0.26	0.18	0.27	0.18	0.26	0.18
	Qin-NR	0.08	0.10	0.08	0.10	0.08	0.12
	P	0.05	0.08	0.13	0.12	0.08	0.09

4.2 Robustness

To investigate the robustness of the proposed estimator, we generate data from model (14) under normal distribution, and consider two methods to create outliers.

C1. We randomly choose four data points, and perturb the covariate x_{ij} to $x_{ij} - 2$ and their responses y_{ij} are replaced by $y_{ij} + 2$;

C2. Eight data points are randomly chosen, and the way to perturb is the same as in C1.

Tabel 3 Simulation results of mean square errors for β under normal distribution (no contamination cases)

		Inde.			Exch.			AR(1)		
Methods		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$n = 50$	He-R	0.0056	0.0064	0.0079	0.0084	0.0056	0.0052	0.0074	0.0043	0.0046
	He-NR	0.0055	0.0064	0.0078	0.0082	0.0051	0.0049	0.0071	0.0042	0.0047
	Qin-R	0.0064	0.0064	0.0074	0.0084	0.0043	0.0059	0.0079	0.0047	0.0059
	Qin-NR	0.0060	0.0061	0.0070	0.0083	0.0046	0.0054	0.0072	0.0047	0.0059
	P	0.0037	0.0052	0.0058	0.0064	0.0031	0.0048	0.0069	0.0035	0.0032
$n = 100$	He-R	0.0027	0.0028	0.0030	0.0035	0.0025	0.0028	0.0046	0.0029	0.0023
	He-NR	0.0028	0.0028	0.0030	0.0036	0.0022	0.0027	0.0046	0.0028	0.0023
	Qin-R	0.0028	0.0020	0.0031	0.0035	0.0021	0.0025	0.0043	0.0027	0.0026
	Qin-NR	0.0027	0.0020	0.0029	0.0037	0.0020	0.0022	0.0044	0.0027	0.0024
	P	0.0025	0.0024	0.0028	0.0030	0.0016	0.0019	0.0041	0.0019	0.0022

Tabel 4 Simulation results of mean square errors for β under contamination cases, where $n = 100$

		Inde.			Exch.			AR(1)		
Methods		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
C1	He-R	0.0053	0.0069	0.0069	0.0090	0.2256	0.0057	0.0082	0.0058	0.0062
	He-NR	0.0129	0.0172	0.0164	0.0179	0.0172	0.0165	0.0171	0.1059	0.0126
	Qin-R	0.0042	0.0070	0.0059	0.0089	0.0054	0.0046	0.0079	0.0055	0.0048
	Qin-NR	0.0081	0.0181	0.0163	0.0189	0.0164	0.0147	0.0157	0.0168	0.0152
	P	0.0038	0.0040	0.0053	0.0058	0.0061	0.0046	0.0064	0.0056	0.0059
C2	He-R	0.0071	0.0103	0.0107	0.0073	0.0091	0.0094	0.0088	0.0103	0.0099
	He-NR	0.0273	0.0358	0.0353	0.0279	0.0396	0.0388	0.0239	0.0348	0.0404
	Qin-R	0.0055	0.0095	0.0109	0.0108	0.0080	0.0091	0.0089	0.0088	0.0101
	Qin-NR	0.0149	0.0404	0.0431	0.0275	0.0395	0.0416	0.0231	0.0399	0.0428
	P	0.0045	0.0091	0.0103	0.0070	0.0087	0.0105	0.0086	0.0069	0.0089

Table 4 and Table 5 show the simulation results for these experiments. It can be seen from these tables, when data is contaminated by outliers, the mean square errors for all estimator increase, especially for He-NR and Qin-NR. For the covariance matrix estimation, the estimator of Qin-NR grows considerably large for QL, while the proposed estimator had a better outlier resistance. In the case of outliers, it can be seen that He-R

Tabele 5 Simulation results for Σ under contamination cases, where $n = 100$

Methods		Inde.		Exch.		AR(1)	
		EL	QL	EL	QL	EL	QL
C1	Qin-R	0.24	0.16	0.38	0.21	0.32	0.20
	Qin-NR	0.81	0.21	1.72	0.36	1.57	0.35
	P	0.19	0.10	0.34	0.26	0.42	0.22
C2	Qin-R	0.33	0.17	0.58	0.26	0.55	0.27
	Qin-NR	2.43	0.38	6.05	0.76	5.52	0.75
	P	0.18	0.09	0.61	0.38	0.54	0.28

and Qin-R gain a lot from using the robust estimate method. Compared with the best performed estimator in each case, the proposed estimator always show comparable outlier resistance.

In conclusion, compared with the existing robust estimators, the proposed estimator show relatively high efficiency under normality, and has comparable resistance to contaminated data set. For the robust estimating equation based estimator, such as He-R and Qin-R, the constant c in the Huber's function $\psi_c(\cdot)$ and the constant r in the weight function can tune the efficiency and robustness of the resulted estimator, and they need be empirically specified in advance, different choice for these constant may show different performance of the estimator. Unlike these methods, the proposed estimator is a weighted least square estimator, and the weights are data driven, since the weights can estimated directly base on generalized empirical likelihood method.

§5. Real Data Analysis

In this section, we applied the proposed estimate methods to analyze the longitudinal CD4 cell count data. A complete description of the data can be found on Diggle P. J.'s homepage: http://www.lancs.ac.uk/diggle/. The data used here is a subset of the complete data and consists of 240 CD4 measurements from first 60 subjects available. The root transformed CD4 counts are taken to be responses and the covariates include years since seroconversion (T), age relative to arbitrary origin (A), smoking status by packs of cigarettes (S1), recreational drug use yes/no (D), number of sex partners (S2), and depression status by the mental illness score (S3). Many authors have analyzed this data set. Wang et al.[8] fitted semi-parametric model and semi-parametric partial linear model for this data and the covariate T entered the models nonparametrically. Qin and Zhu[21] also fitted a partial linear mixed model for this data and give the robust estimation of the parameters in the model. Let y_{ij} denote the jth CD4 counts value measured for the ith subject, and by T_{ij}, A_{ij}, $\mathrm{S1}_{ij}$, D_{ij}, $\mathrm{S2}_{ij}$, $\mathrm{S3}_{ij}$ the corresponding covariates values. We standardize these data, and consider the following linear model:

$$y_{ij} = \beta_0 + \beta_1 \mathrm{T}_i + \beta_2 \mathrm{T}_i^2 + \beta_3 \mathrm{A}_{ij} + \beta_4 \mathrm{S1}_{ij} + \beta_5 \mathrm{D}_{ij} + \beta_6 \mathrm{S2}_{ij} + \beta_7 \mathrm{S3}_{ij} + \epsilon_{ij}, \tag{17}$$

where we include the T^2 in the model to incorporate possible nonlinear relationship between T and response value. Then, we apply the methods used in simulation study to model (17), and summarize the results in Table 6, where the results for the proposed estimator is based on 1 000 bootstrap sample of CD4 data set. It is clear that the proposed estimator is similar to the robust estimators Qin-R and He-R, since these estimators show the significance for covariate T and S1. Furthermore, T show positive effects on the CD4 numbers while S1 present negative effects. Meanwhile, non-robust estimator Qin-NR and He-NR differ with the rest estimator on the significance for S3. This difference indicates the influence induced by possible outliers in CD4 data. We apply the algorithm in section 3 to CD4 data, and compute the weights for the observations to identify possible outliers. It can be seen from following Figure 1, the 32th, 58th, 150th, 198th, 225th observations are heavily down-weighted. These means these observations are outliers of CD4 cell count data under linear regression model.

Tabel 6 Coefficient estimates for CD4 data, with standard errors in parentheses. We use the exchangeable working correlation structure for He-R and He-NR methods. We show the significance (two side) at level 0.05 with $*$, or at level 0.01 with $**$

	Intercept	T	T^2	A	S1	D	S2	S3
He-R	-0.0506	-0.2715**	-0.0264	-0.0990	0.2944**	0.0920	0.1276	-0.0784
	(0.0712)	(0.0724)	(0.0611)	(0.0675)	(0.0743)	(0.0713)	(0.0669)	(0.0508)
He-NR	-0.0511	-0.2761**	-0.0261	-0.1121	0.2635**	0.1109	0.1324	-0.1181*
	(0.0593)	(0.0608)	(0.0714)	(0.0694)	(0.0822)	(0.0749)	(0.0752)	(0.0520)
Qin-R	-0.0484	-0.3056**	-0.0279	-0.0585	0.3173**	0.0233	0.0313	-0.0691
	(0.0670)	(0.0684)	(0.0595)	(0.0738)	(0.0775)	(0.0688)	(0.0667)	(0.0523)
Qin-NR	-0.0660	-0.2936**	-0.0144	-0.0792	0.3143**	0.0103	0.0475	-0.1167*
	(0.0702)	(0.0785)	(0.0656)	(0.0729)	(0.0784)	(0.0673)	(0.0641)	(0.0582)
P	-0.0419	-0.2881**	-0.0408	-0.0786	0.3005**	0.0963	0.1120	-0.0981
	(0.0673)	(0.0654)	(0.05801)	(0.0638)	(0.0532)	(0.0746)	(0.0733)	(0.0788)

We give the estimate for the correlation matrix of CD4 data set in Table 7. The correlation matrix estimate for the proposed method is also base on 1 000 bootstrap sample and the mean of resulted covariance matrix estimates. Compared with non-robust estimate

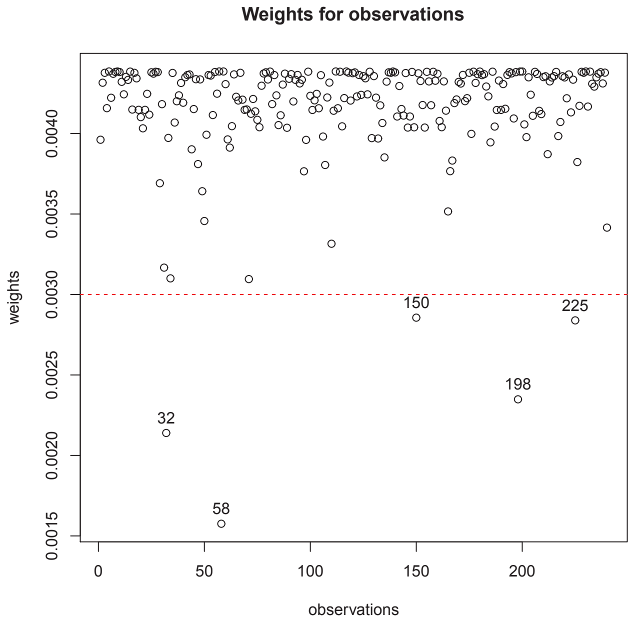

Figure 1 240 observation labeled on horizontal axis and their weights labeled on vertical axis

method Qin-NR, the estimated correlation matrix by the proposed method and Qin-R show relatively higher correlation among the four observations within each subjects.

Table 7 Correlation matrix estimates in the analysis of the CD4 data

Qin-R				Qin-NR				P			
1.0000	0.9749	0.4914	0.5405	1.0000	0.5378	0.3548	0.2527	1.0000	0.7906	0.4059	0.4966
0.9749	1.0000	0.6598	0.5823	0.5378	1.0000	0.4186	0.3679	0.7906	1.0000	0.4275	0.6014
0.4914	0.6598	1.0000	0.5427	0.3584	0.4186	1.0000	0.2432	0.4059	0.4275	1.0000	0.3966
0.5405	0.5823	0.5427	1.0000	0.2527	0.3679	0.2432	1.0000	0.4966	0.6014	0.3966	1.0000

§6. Conclusion and Further Work

In this paper, we propose an efficient robust estimate for mean and covariance for longitudinal data in regression models, which is more efficient under normal case, and also enjoy comparable outlier resistance with existing robust estimator. The impressive finite sample performance of the proposed estimator shows the power of generalized empirical likelihood type estimator for tailoring estimators to achieve specific objectives, and this may be due to the second order behavior of saddlepoint based procedures, see [24].

The breakdown properties and large sample variance estimation still require further careful investigation. Bondell and Stefanski[19] studied the breakdown properties and asymptotic normality for independent data. However, it may be more complicated here since we study the estimator of mean and covariance simultaneously based on Cholesky decomposition, and the variation induced by the estimator $\widehat{\epsilon}$ has to be considered.

To further enhance the estimator's robustness is another interesting topic. A possible way to achieve this goal is to consider additional constraints to bound influence and leverage in (10), but this remain an open question and a possible line of future work.

References

[1] LIANG K Y, ZEGER S L. Longitudinal data analysis using generalized linear models [J]. *Biometrika*, 1986, **73(1)**: 13–22.

[2] QU P Y, LINDSAY B G, Bing L I. Improving generalised estimating equations using quadratic inference functions [J]. *Biometrika*, 2000, **87(4)**: 823–836.

[3] YE H J, PAN J X. Modelling of covariance structures in generalised estimating equations for longitudinal data [J]. *Biometrika*, 2006, **93(4)**: 927–941.

[4] WU W B, POURAHMADI M. Nonparametric estimation of large covariance matrices of longitudinal data [J]. *Biometrika*, 2003, **90(4)**: 831–844.

[5] PAN J X, MACKENZIE G. On modelling mean-covariance structures in longitudinal studies [J]. *Biometrika*, 2003, **90(1)**: 239–244.

[6] FAN J Q, WU Y C. Semiparametric estimation of covariance matrices for longitudinal data [J]. *J Amer Statist Assoc*, 2008, **103(484)**: 1520–1533.

[7] LENG C L, ZHANG W P, PAN J X. Semiparametric mean-covariance regression analysis for longitudinal data [J]. *J Amer Statist Assoc*, 2010, **105(489)**: 181–193.

[8] WANG N, CARROLL R J, LIN X H. Efficient semiparametric marginal estimation for longitudinal/clustered data [J]. *J Amer Statist Assoc*, 2005, **100(469)**: 147–157.

[9] CANTONI E. A robust approach to longitudinal data analysis [J]. *Canad J Statist*, 2004, **32(2)**: 169–180.

[10] HE X M, FUNG W K, ZHU Z Y. Robust estimation in generalized partial linear models for clustered data [J]. *J Amer Statist Assoc*, 2005, **100(472)**: 1176–1184.

[11] QIN G Y, ZHU Z Y. Robust estimation in generalized semiparametric mixed models for longitudinal data [J]. *J Multivariate Anal*, 2007, **98(8)**: 1658–1683.

[12] ZHENG X Y, FUNG W K, ZHU Z Y. Robust estimation in joint mean-covariance regression model for longitudinal data [J]. *Ann Inst Statist Math*, 2013, **65(4)**: 617–638.

[13] QIN G Y, ZHU Z Y. Robust estimation of mean and covariance for longitudinal data with dropouts [J]. *J Appl Stat*, 2015, **42(6)**: 1240–1254.

[14] OWEN A B. Empirical likelihood ratio confidence intervals for a single functional [J]. *Biometrika*, 1988, **75(2)**: 237–249.

[15] YOU J H, CHEN G M, ZHOU Y. Block empirical likelihood for longitudinal partially linear regression models [J]. *Canad J Statist*, 2006, **34(1)**: 79–96.

[16] XUE L G, ZHU L X. Empirical likelihood semiparametric regression analysis for longitudinal data [J]. *Biometrika*, 2007, **94(4)**: 921–937.

[17] WANG S J, QIAN L F, CARROLL R J. Generalized empirical likelihood methods for analyzing longitudinal data [J]. *Biometrika*, 2010, **97(1)**: 79–93.

[18] QIN J, LAWLESS J. Empirical likelihood and general estimating equations [J]. *Ann Statist*, 1994, **22(1)**: 300–325.

[19] BONDELL H D, STEFANSKI L A. Efficient robust regression via two-stage generalized empirical likelihood [J]. *J Amer Statist Assoc*, 2013, **108(502)**: 644–655.

[20] SINHA S K. Robust analysis of generalized linear mixed models [J]. *J Amer Statist Assoc*, 2004, **99(466)**: 451–460.

[21] QIN G Y, ZHU Z Y. Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data [J]. *Biometrics*, 2009, **65(1)**: 52–59.

[22] CRESSIE N, READ T R C. Multinomial goodness-of-fit tests [J]. *J Roy Statist Soc Ser B*, 1984, **46(3)**: 440–464.

[23] CORCORAN S A. Bartlett adjustment of empirical discrepancy statistics [J]. *Biometrika*, 1998, **85(4)**: 967–972.

[24] IMBENS G W. Generalized method of moments and empirical likelihood [J]. *J Bus Econom Statist*, 2002, **20(4)**: 493–506.

[25] ROUSSEEUW P J, LEROY A M. *Robust Regression and Outlier Detection* [M]. New York: Wiley, 1987.

[26] HE X M, PORTNOY S. Reweighted LS estimators converge at the same rate as the initial estimator [J]. *Ann Statist*, 1992, **20(4)**: 2161–2167.

[27] OWEN A B. *Empirical Likelihood* [M]. New York: Chapman and Hall/CRC, 2001.

[28] WANG N. Marginal nonparametric kernel regression accounting for within-subject correlation [J]. *Biometrika*, 2003, **90(1)**: 43–52.

纵向数据均值和协方差矩阵的有效稳健估计

樊亚莉　　徐孝琳

(上海理工大学理学院, 上海, 200093)

摘　要: 本文提出一种针对纵向数据回归模型下的均值和协方差矩阵同时进行的有效稳健估计. 基于对协方差矩阵的 Cholesky 分解和对模型的改写, 我们提出一个加权最小二乘估计, 其中权重是通过广义经验似然方法估计出来的. 所提估计的有效性得益于经验似然方法的优势, 稳健性则是通过限制残差平方和的上界来达到. 模拟研究表明, 和已有的针对纵向数据的稳健估计相比, 所提估计具有更高的效率和可比的稳健性. 最后, 我们把所提估计方法用来分析一组实际数据.

关键词: 有效估计; 广义经验似然; 稳健性

中图分类号: O212.1