VARIABLE SELECTION IN ROBUST JOINT MEAN AND COVARIANCE MODEL FOR LONGITUDINAL DATA ANALYSIS

Xueying Zheng, Wing Kam Fung and Zhongyi Zhu

Supplementary Material

S1 Sketch of proofs

Following Johnson, Lin, and Zeng (2008), we employ a zero-crossing estimator $\hat{\theta}$ to the penalized estimating equation if, for $j = 1, \ldots, s$,

$$\overline{\lim_{\varepsilon\to 0^+}} \frac{1}{m} U_j(\hat{\theta}+\varepsilon e_j)U_j(\hat{\theta}-\varepsilon e_j) \leq 0,$$

where U_j is the jth component of $U(\cdot)$ and e_j is the jth canonical unit vector. Furthermore, an approximate zero-crossing estimator $\hat{\theta}$ is defined if

$$\overline{\lim_{m\to\infty}}\ \overline{\lim_{\varepsilon\to 0^+}} \frac{1}{m} U_j(\hat{\theta}+\varepsilon e_j)U_j(\hat{\theta}-\varepsilon e_j) \leq 0.$$

It is obvious that when U is continuous, the zero-crossing estimator is the same as an exact solution to the penalized estimating equation. Different from only the mean penalized estimating equation in Johnson , Lin, and Zeng (2008), the proposed approximate zero-crossing estimator is for both the mean and the covariance robust penalized estimating equations.

Denote

$$U^R(\theta) = ([U_1^R(\beta)]^T,\ [U_2^R(\gamma)]^T,\ [U_3^R(\lambda)]^T)^T,$$

where

$$U_1^R(\beta) = \sum_{i=1}^m X_i^T (V_i^\beta)^{-1} h_i^\beta(\mu_i(\beta)), \tag{S1.1}$$

$$U_2^R(\gamma) = \sum_{i=1}^m T_i^T (V_i^\gamma)^{-1} h_i^\gamma(\hat{r}_i(\gamma)), \tag{S1.2}$$

$$U_3^R(\lambda) = \sum_{i=1}^m Z_i^T D_i (V_i^\lambda)^{-1} h_i^\lambda(\sigma_i^2(\lambda)). \tag{S1.3}$$

The following assumptions are required to establish asymptotic properties.

(C.1) The covariate vectors are fixed. Also, for each subject the number of repeated measurements, n_i, is fixed.

(C.2) The design matrices in the joint models are all bounded, meaning that all the elements of the matrices are bounded by a single finite real number. The first four moments of y_{ij} exist.

(C.3) The random functions $U^R(\cdot)$ satisfy Lipschitz condition, that is, there exists a measurable function $m(\cdot)$ such that, $||U^R(\theta_1,\cdot) - U^R(\theta_2,\cdot)|| \le m(\cdot)||\theta_1 - \theta_2||$, for every θ_1, θ_2 in the neighborhood of θ_0, where $\int m^2(\cdot)dP < \infty$.

(C.4) For some $\delta > 0$, we assume

$$\sup_i\{E||h_{0i}^{\beta}(\mu_i)||^{2+\delta}, E||h_{0i}^{\gamma}(\hat{r}_i)||^{2+\delta}, E||h_{0i}^{\lambda}(\sigma_i^2)||^{2+\delta}\} < \infty,$$

where $h_{0i}(\cdot)$ is similar to $h_i(\cdot)$ but the former is evaluated at the true value of parameters. Moreover,

$$Eh_{0i}^{\beta}(\mu_i)h_{0i}^{\beta}(\mu_i)^T = F_i^{\beta} > 0,\ Eh_{0i}^{\gamma}(\hat{r}_i)h_{0i}^{\gamma}(\hat{r}_i)^T = F_i^{\gamma} > 0,\ Eh_{0i}^{\lambda}(\sigma_i^2)h_{0i}^{\lambda}(\sigma_i^2)^T = F_i^{\lambda} > 0$$

with $\sup_i\{||F_i^{\beta}||, ||F_i^{\gamma}||, ||F_i^{\lambda}||\} < \infty$.

(C.5) The function $C_i^{\beta}(\mu_i) = E[\psi(A_i^{-1/2}(y_i - \mu_i))]$, $C_i^{\gamma}(\hat{r}_i) = E[\psi(D_i^{-1/2}(r_i - \hat{r}_i))]$ and $C_i^{\lambda}(\sigma_i^2) = E[\psi(\widetilde{A}_i^{-1/2}(\varepsilon_i^2 - \sigma_i^2))]$ has bounded second derivative. The functions $\psi(\cdot)$ is piecewise twice differentiable, and the second derivatives are bounded.

(C.6) (a) For SCAD penalty we assume the tuning parameters τ_m^{β}, τ_m^{γ} and τ_m^{λ} satisfying $\tau_m \to 0$, $\sqrt{m}\tau_m \to \infty$ when $m \to \infty$.

(b) For ALASSO penalty we assume the tuning parameters τ_m^{β}, τ_m^{γ} and τ_m^{λ} satisfying $\sqrt{m}\tau_m \to 0$, $m\tau_m \to \infty$ when $m \to \infty$.

By (C.3), the convergence is uniform about θ in the neighborhood of θ_0. (C.4) and (C.5) imposed on the score function ψ can be easily checked to be satisfied when ψ is bounded as that in section 2.2. (C.6) guarantees the oracle properties when we use SCAD or ALASSO penalty.

Proof of Theorem 1.

The proof is a generalization of that in Johnson, Lin, and Zeng (2008). On one hand, we propose three estimating equations for parameters in the mean regression model together with the decomposed covariance matrices. Moreover, we robustify the three penalized estimating equations.

To prove (a) in Theorem 1, we verify that $\hat{\theta} = (\hat{\theta}^{s_1 T}, 0^T)^T$ is an approximate zero-crossing estimator for $U(\cdot)$, where $\hat{\theta}^{s_1} = \theta_0^{s_1} - m^{-1}(G^{s_1})^{-1}U_{s_1}^R(\theta_0)$. Here $U_{s_1}(\cdot)$ and $U_{s_1}^R(\cdot)$ denote the nonzero s_1-components of $U(\cdot)$ and $U^R(\cdot)$.

The following result

$$\sup_{||\theta-\theta_0||<Mm^{-1/2}} ||m^{-1/2}U^R(\theta) - m^{-1/2}U^R(\theta_0) - m^{1/2}G_m(\theta - \theta_0)|| = o_p(1), \qquad \text{(S1.4)}$$

was assumed in Johnson, Lin, and Zeng (2008). It can also be obtained by our conditions (C.3) and (C.4), because (C.3) indicates that $\{U(\theta) : \theta \in \Theta, \Theta \subseteq R^{p+q+d}\}$ is a Donsker class of empirical process and the convergence is uniform about θ in the neighborhood of θ_0.

Meanwhile, conditions for multivariate Lyapunov central limit theorem for U_θ^R can be verified by the conditions (C.3)-(C.5) we imposed, i.e, $m^{-1/2}(G^{s_1})^{-1}U_{s_1}^R(\theta_0) \to N_{s_1}(0,\ (G^{s_1})^{-1}B^{s_1}[(G^{s_1})^T]^{-1})$, as $n \to \infty$ in distribution.

Therefore under (S1.4) and $\hat{\theta} = \theta_0 + O_p(m^{-1/2})$, we have

$$m^{-1/2}U_j(\hat{\theta} \pm \varepsilon e_j) = o_p(1) - m^{-1/2}q_{\tau_m}(|\hat{\theta}_j \pm \varepsilon|)\mathrm{sgn}(|\hat{\theta}_j \pm \varepsilon|), \quad j = 1, \ldots, s_1,$$

where $U_j(\cdot)$ is the jth component of $U(\cdot)$.

Under condition (a) in (C.6), for any nonzero fixed $\theta \in \Theta$, we have $\lim_{m\to\infty} \sqrt{m}q_{\tau_m}(|\theta|) = \lim_{m\to\infty} q'_{\tau_m}(|\theta|) = 0$ for SCAD penalty. For ALASSO penalty, given condition (b) in (C.6) we conclude $\sqrt{m}q_{\tau_m}(|\theta|) = \sqrt{m}\tau_m/|\tilde{\theta}| \to 0$ if $1/|\tilde{\theta}| < \infty$ and $q'_{\tau_m}(|\theta|) = 0$, where $\tilde{\theta}$ stands for the regression coefficient estimates obtained from $U^R(\theta)$ (without penalty). As a consequence, for any $\varepsilon \to 0^+$, $m^{-1/2}U_j(\hat{\theta} \pm \varepsilon e_j) = o_p(1)$, $j = 1, \ldots, s_1$.

Furthermore, for SCAD and any $M > 0$, $\sqrt{m}\inf_{|\theta|\le Mm^{-1/2}} q_{\tau_m}(|\theta|) = \sqrt{m}\tau_m \to \infty$. When $\sqrt{m}(\tilde{\theta} - \theta_0) = O_p(1)$ under some conditions, then $\sqrt{m}\inf_{|\theta|\le Mm^{-1/2}} q_{\tau_m}/|\tilde{\theta}| = M^{-1}m\tau_m \to \infty$ for ALASSO. Therefore, for $j = s_1+1, \ldots, s$, $\hat{\theta}_j = 0$, $m^{-1/2}U_j(\hat{\theta} + \varepsilon e_j)$ and $m^{-1/2}U_j(\hat{\theta} - \varepsilon e_j)$ are dominated by $-m^{-1/2}q_{\tau_m}(\varepsilon)$ and $m^{-1/2}q_{\tau_m}(\varepsilon)$ with opposite signs when $\varepsilon \to 0$. As a result, we conclude that $\hat{\theta}$ is an approximate zero-crossing estimator by its definition.

We denote $\hat{\theta} = (\hat{\theta}_1, \ldots \hat{\theta}_s)^T$ as $\sqrt{m}$ consistent approximate zero-crossing solution of $U(\theta) = 0$. To prove (b) in Theorem 1, we need to show that for any $\varepsilon > 0$, when m is sufficiently large,

$$P\{\hat{\theta}_j \neq 0, j = s_1+1, \ldots, s\} < \varepsilon.$$

On one hand, for $j = s_1+1, \ldots, s$, there exists some $M > 0$ such that when m is large enough, $P\{\hat{\theta}_j \neq 0, j = s_1+1, \ldots, s, |\hat{\theta}_j| \ge Mm^{-1/2}\} < \varepsilon/2$. Therefore, we only need to show that $P\{\hat{\theta}_j \neq 0, j = s_1+1, \ldots, s, |\hat{\theta}_j| < Mm^{-1/2}\} < \varepsilon/2$ when m is large enough.

As shown in Johnson, Lin, and Zeng (2008), $[m^{-1/2}U_j(\hat{\theta})]^2 = o_p(1)$ and there exists some $M' > 0$ such that for large m,

$$P\{\hat{\theta}_j \neq 0, j = s_1+1, \ldots, s, |\hat{\theta}_j| < Mm^{-1/2},\ m^{1/2}q_{\tau_m}(|\hat{\theta}_j|) > M'\} < \varepsilon/2.$$

Because for any $M > 0$, $\lim_{m\to\infty} \sqrt{m}\inf_{|\theta|\le Mm^{-1/2}} q_{\tau_m}(|\theta|) \to \infty$, by condition (C.6), $\hat{\theta}_j \neq 0$, $j = s_1+1, \ldots, s$, $|\hat{\theta}|_j < Mm^{-1/2}$ implies that $m^{-1/2}q_{\tau_m}(|\hat{\theta}_j|) > M'$ for large m. Thus $P\{\hat{\theta}_j \neq 0, j = s_1+1, \ldots, s, |\hat{\theta}_j| < Mm^{-1/2}\} < \varepsilon/2$ and Theorem 1 (b) has been proved.

Proof of Theorem 2.

We apply the Taylor series expansion on the last term of

$$m^{-1/2}U_{s_1}^R(\theta_0) + m^{1/2}G^{s_1}(\hat{\theta}^{s_1} - \theta_0^{s_1}) - m^{1/2}q_{\tau_m}(|\hat{\theta}^{s_1}|)\mathrm{sgn}(\hat{\theta}^{s_1}) = o_p(1),$$

and rearranged by $c_m = (q_{\tau_m}(|\theta_{01}^{s_1}|)\mathrm{sgn}(\theta_{01}^{s_1}), \ldots, q_{\tau_m}(|\theta_{0s_1}^{s_1}|)\mathrm{sgn}(\theta_{0s_1}^{s_1}))^T$ and $\Omega_m^{s_1}$ to obtain that

$$\sqrt{m}(G_m^{s_1} + \Omega_m^{s_1})\{\hat{\theta}_m^{s_1} - \theta_0^{s_1} + (G_m^{s_1} + \Omega_m^{s_1})^{-1}c_m\} \to_d N_{s_1}(0, B^{s_1}).$$

For any nonzero fixed $\theta \in \Theta$, under conditions (a) and (b) in (C.6), $\lim_{m\to\infty}\sqrt{m}q_{\tau_m}(|\theta|) = \lim_{m\to\infty} q'_{\tau_m}(|\theta|) = 0$ for SCAD penalty and $\sqrt{m}q_{\tau_m}(|\theta|) = \sqrt{m}\tau_m/|\tilde{\theta}| \to 0$ if $1/|\tilde{\theta}| < \infty$ and $q'_{\tau_m}(|\theta|) = 0$ for ALASSO penalty, where $\tilde{\theta}$ stands for the regression coefficient estimates obtained from $U^R(\theta)$ (without penalty). Therefore, c_m tends to zero for both SCAD and ALASSO penalized estimators as $m \to \infty$.

S2 An algorithm

We describe the algorithm for estimating β, γ and λ simultaneously. The iterative MM algorithm can be written as:

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} + [G(\beta^{(k)}) + \Delta_{\tau^{(1)}}(\beta^{(k)})]^{-1}U_1(\beta^{(k)}),\\
\gamma^{(k+1)} &= \gamma^{(k)} + [G(\gamma^{(k)}) + \Delta_{\tau^{(2)}}(\gamma^{(k)})]^{-1}U_2(\gamma^{(k)}),\\
\lambda^{(k+1)} &= \lambda^{(k)} + [G(\lambda^{(k)}) + \Delta_{\tau^{(3)}}(\lambda^{(k)})]^{-1}U_3(\lambda^{(k)}),
\end{aligned}$$

where $G(\beta)$, $G(\gamma)$ and $G(\lambda)$ are $p \times p$, $q \times q$ and $d \times d$ submatrices of G corresponding to the parameters, and

$$\Delta_{\tau^{(1)}}(\beta) = \mathrm{diag}(\frac{q_{\tau^{(1)}}(|\beta_1|)}{\varepsilon_p + |\beta_1|}, \cdots, \frac{q_{\tau^{(1)}}(|\beta_p|)}{\varepsilon_p + |\beta_p|}),$$

$$\Delta_{\tau^{(2)}}(\gamma) = \mathrm{diag}(\frac{q_{\tau^{(2)}}(|\gamma_1|)}{\varepsilon_p + |\gamma_1|}, \cdots, \frac{q_{\tau^{(2)}}(|\gamma_q|)}{\varepsilon_p + |\gamma_q|}),\ \Delta_{\tau^{(3)}}(\lambda) = \mathrm{diag}(\frac{q_{\tau^{(3)}}(|\lambda_1|)}{\varepsilon_p + |\lambda_1|}, \cdots, \frac{q_{\tau^{(3)}}(|\lambda_d|)}{\varepsilon_p + |\lambda_d|}).$$

Here ε_p is a fixed small number for the purpose of preventing computational singular, and chosen to be 10^{-6} in our simulation.

Convenient initial values for the covariance parameters are $\gamma = 0$ and $\lambda = 0$, which indicate independence structure in the covariance matrix. A natural choice of the initial estimate $\beta^{(0)}$ of β is the solution to (2.2) in this special case as the robust GEE estimator under working independence covariance structure, whose consistency has been proved. The following algorithm summarizes the computation of the penalized GEE estimators for joint mean and covariance model.
Algorithm:

0. Take zeros as initial values of γ and λ, that is $\gamma^{(0)}$ and $\lambda^{(0)}$. Then take $\beta^{(0)}$ as the the solution to (2.2) with independent working matrix.

1. Given current $\{\beta^{(k)},\ \gamma^{(k)},\ \lambda^{(k)}\}$, use the MM algorithm to update γ and λ until convergence, which are denoted as $\gamma^{(k+1)}$ and $\lambda^{(k+1)}$.

2. For the updated $\gamma^{(k+1)}$ and $\lambda^{(k+1)}$, form

$$\phi_{ijk}^{(k+1)} = g_{ijk}^T \gamma^{(k+1)}, \text{ and } (\sigma_{ij}^2)^{(k+1)} = \exp(z_{ij}^T \lambda^{(k+1)})$$

to construct

$$\Sigma_i^{(k+1)} = (\Phi_i^{(k+1)})^{-1} D_i^{(k+1)} [(\Phi_i^{(k+1)})^T]^{-1}.$$

Then update β according to the MM algorithm to obtain $\beta^{(k+1)}$.

3. Repeat Step 1 and Step 2 above until convergence.

This widely-used algorithm (the non-robust version, i.e. the tuning parameter of Huber function c is set to a very large value) has been proved to be well performed under the usual case (no outliers) by Kou and Pan (2011). However, in our simulation, we find the non-robust algorithm behaves poorly when there are outliers. Since the joint mean and covariance model introduces more parameters than GEE, the algorithm based on the non-robust penalized estimating equations faces a great challenge to obtain a reliable estimator when the contamination can happen in any of three estimating equations. Even a tiny contamination can result in non-convergence or estimation with a large bias. In contrast, the robust algorithm performs fairly well against perturbations.

S3 Standard errors and losses of covariance matrix estimators in Studies 1 and 2

To judge the performance of the proposed covariance estimator, we introduce two commonly used loss functions: the entropy loss

$$L_1(\Sigma, \hat{\Sigma}) = m^{-1} \sum_{i=1}^{m} \{\mathrm{trace}(\Sigma_i \hat{\Sigma}_i^{-1}) - \log |\Sigma \hat{\Sigma}_i^{-1}| - n_i\}$$

and the quadratic loss

$$L_2(\Sigma, \hat{\Sigma}) = m^{-1} \sum_{i=1}^{m} \mathrm{trace}(\Sigma_i^{-1} \hat{\Sigma}_i - I_i)^2,$$

where Σ_i and $\hat{\Sigma}_i$ are the true covariance matrix and its estimator.

Table S3.1: Standard errors for SCAD estimators in Study 1

	$\gamma = 0$				$\gamma \neq 0$			
	n=100		n=200		n=100		n=200	
	NR	R	NR	R	NR	R	NR	R
				NC				
β_1	0.023	0.025	0.016	0.018	0.022	0.026	0.016	0.018
β_4	0.027	0.030	0.018	0.020	0.025	0.031	0.016	0.024
β_5	0.028	0.031	0.019	0.022	0.025	0.030	0.016	0.020
β_{10}	0.027	0.029	0.017	0.020	0.026	0.030	0.015	0.018
γ_1	0.006	0.006	0.002	0.003	0.008	0.009	0.005	0.006
λ_2	0.095	0.109	0.064	0.073	0.087	0.101	0.063	0.072
λ_6	0.076	0.085	0.052	0.058	0.074	0.089	0.052	0.064
				C0				
β_1	0.052	0.029	0.041	0.019	0.056	0.043	0.044	0.023
β_4	0.038	0.030	0.037	0.020	0.050	0.043	0.053	0.024
β_5	0.043	0.034	0.036	0.023	0.049	0.038	0.033	0.028
β_{10}	0.036	0.028	0.030	0.019	0.045	0.038	0.029	0.022
γ_1	0.014	0.007	0.004	0.002	0.017	0.014	0.016	0.002
λ_2	0.152	0.112	0.158	0.074	0.182	0.122	0.222	0.078
λ_6	0.229	0.092	0.194	0.058	0.294	0.119	0.265	0.064

Table S3.2: Standard errors for mean estimators in Study 2

	NC					C1				
	rpj$_{ar}$	rpj$_{in}$	rpgee$_{ar}$	rpgee$_{ex}$	rpgee$_{in}$	rpj$_{ar}$	rpj$_{in}$	rpgee$_{ar}$	rpgee$_{ex}$	rpgee$_{in}$
						IN				
β_1	0.048	0.048	0.047	0.047	0.047	0.054	0.054	0.057	0.056	0.057
β_4	0.049	0.049	0.047	0.047	0.047	0.050	0.050	0.052	0.052	0.052
β_5	0.051	0.051	0.051	0.051	0.051	0.056	0.056	0.059	0.059	0.059
β_{10}	0.048	0.048	0.047	0.047	0.047	0.051	0.051	0.061	0.061	0.061
						AR				
β_1	0.055	0.055	0.046	0.050	0.054	0.059	0.059	0.059	0.059	0.063
β_4	0.056	0.056	0.046	0.048	0.054	0.063	0.063	0.062	0.061	0.064
β_5	0.061	0.061	0.050	0.057	0.060	0.067	0.067	0.065	0.066	0.067
β_{10}	0.054	0.054	0.046	0.052	0.054	0.057	0.057	0.063	0.065	0.066
						EX				
β_1	0.056	0.056	0.043	0.041	0.057	0.061	0.061	0.061	0.058	0.065
β_4	0.064	0.064	0.052	0.045	0.063	0.069	0.069	0.066	0.062	0.069
β_5	0.061	0.061	0.050	0.048	0.060	0.068	0.068	0.063	0.060	0.066
β_{10}	0.056	0.056	0.045	0.044	0.056	0.060	0.060	0.062	0.060	0.066

Three normal covariance structures are set as the truth: working independence (IN), auto-regressive (AR) and exchangeable (EX) with correlation parameter 0.5. rpj$_{ar}$ and rpj$_{in}$ are robust joint models with independent and AR(1) correlation structure in (2.4,) respectively; rpgeee$_{in}$, rpgeee$_{ar}$, and rpgeee$_{ex}$ are robust penalized GEE estimations with IN, AR and EX as working correlation matrices.

Table S3.3: Entropy loss (L_1) and quadratic loss (L_2) in estimating Σ in Study 2

		IN		AR		EX	
		L_1	L_2	L_1	L_2	L_1	L_2
NC	rpj_{ar}	0.052	0.027	1.213	0.043	1.728	0.054
	rpj_{in}	0.052	0.028	1.211	0.043	1.726	0.053
	rpgee_{ar}	0.013	0.014	0.039	0.036	0.570	0.045
	rpgee_{ex}	0.012	0.015	0.509	0.025	0.040	0.038
	rpgee_{in}	0.003	0.004	1.155	0.005	1.679	0.007
	pgee_{ar}	0.015	0.017	0.019	0.022	0.549	0.033
	pgee_{ex}	0.014	0.017	0.505	0.022	0.019	0.023
	pgee_{in}	0.003	0.004	1.155	0.005	1.679	0.007
C1	rpj_{ar}	0.042	0.042	1.205	0.050	1.727	0.064
	rpj_{in}	0.042	0.042	1.206	0.052	1.728	0.065
	rpgee_{ar}	1.079	0.071	1.952	0.088	2.484	0.090
	rpgee_{ex}	1.079	0.071	2.009	0.087	2.278	0.101
	rpgee_{in}	1.080	0.071	2.218	0.072	2.745	0.077
	pgee_{ar}	1.046	0.069	1.877	0.120	2.410	0.120
	pgee_{ex}	1.044	0.069	1.932	0.103	2.168	0.135
	pgee_{in}	1.046	0.069	2.187	0.070	2.713	0.074

S4 Estimates for mean after removing the outlier in hormone data analysis

Table S4.1: Estimates of the mean parameters β and standard errors (inside brackets) after removing the outlier (observation 10 of subject 1)

	Intercept	Age	BMI	Time	Time^2	Age× BMI	Age× Time	BMI× Time
rpj	0.882	0	0	0.717	0	0	0	0
	(0.074)	(-)	(-)	(0.053)	(-)	(0)	(-)	(0)
pj	0.890	0	0	0.684	0	0	0	0
	(0.075)	(-)	(-)	(0.054)	(-)	(-)	(-)	(-)
gee	0.981	-	-	0.746	-	-	-	-
	(0.088)	-	-	(0.045)	-	-	-	-