Spatial Linear Mixed Models with Covariate Measurement Errors

Yi Li, Haicheng Tang and Xihong Lin

Harvard University and American Express

Supplemental Material

S1. Verification of Conditions (c.1) and (c.2) for the Exponential, Guassian and CAR models

To show (c.1) hold, we use the matrix norm property that the spectral radius of any matrix $\mathbf{G}$ is no larger than its row sum norm, denoted by $||\mathbf{G}||_\infty$ (Theorem 5.6.7 of Graybill, 1969). Let λ_n be the largest eigenvalue of $\mathbf{\Lambda} = \mathbf{V}(\theta) + \sigma_\epsilon^2\mathbf{I} = \theta\mathbf{R} + \sigma_\epsilon^2\mathbf{I}$, where $\mathbf{R}$ is the spatial correlation matrix defined in Section 2. Thus

$$\lambda_n \leq \theta||\mathbf{R}||_\infty + \sigma_\epsilon^2.$$

We now study $||\mathbf{R}||_\infty$ under the the exponential model, the Gaussian model and the CAR model, respectively. First, consider the exponential model on a regular grid $[0, \sqrt{n}]^2$,

$$\begin{aligned} ||\mathbf{R}||_\infty &\leq max_{i,j} \sum_{k_1=0}^{\sqrt{n}-1} \sum_{k_2=0}^{\sqrt{n}-1} e^{-\sqrt{(k_1-i)^2+(k_2-j)^2}} \\ &\leq max_{i,j} \sum_{k_1=0}^{\sqrt{n}-1} \sum_{k_2=0}^{\sqrt{n}-1} e^{-(|k_1-i|+|k_2-j|)/\sqrt{2}} \\ &\leq \sum_{k_1=0}^{\sqrt{n}-1} 2e^{-k_1/\sqrt{2}} \sum_{k_2=0}^{\sqrt{n}-1} 2e^{-k_2/\sqrt{2}} \\ &= 4\Big(\frac{1-e^{-\sqrt{n/2}}}{1-e^{-1/\sqrt{2}}}\Big)^2 \\ &< \frac{4}{(1-e^{-1/\sqrt{2}})^2}. \end{aligned} \tag{S1.1}$$

Thus $||\mathbf{R}||_\infty$ is bounded by the constant $\frac{4}{(1-e^{-1/\sqrt{2}})^2}$ for any n. Secondly, with the "Gaussian" spatial correlation, noting

$$||\mathbf{R}||_\infty \leq max_{i,j} \sum_{k_1=0}^{\sqrt{n}-1} \sum_{k_2=0}^{\sqrt{n}-1} e^{-((k_1-i)^2+(k_2-j)^2)} \tag{S1.2}$$

and

$$e^{-(k_1-i)^2+(k_2-j)^2)} \leq e^{-\sqrt{(k_1-i)^2+(k_2-j)^2}} \tag{S1.3}$$

will lead to the same bound as in (S1.1). Finally, consider the CAR model. Denote by $\mathbf{A} = \mathbf{M}^{-1} - \gamma\mathbf{Q} \overset{def}{=} (a_{ij})_{n\times n}$, and denote its eigenvalues by $\delta_i, i = 1, \ldots, n$. Using the Gerschgorin Disc Theorem (see e.g. Wilkinson, 1965), we have that $\tilde{\lambda}_i \in \cup_{j=1}^n \{z \in R : |z - a_{jj}| \leq \sum_{k\neq j} |a_{jk}|\}$. Hence, $\min_j(a_{jj} - \sum_{k\neq j}|a_{jk}|) = \min_j(q_{j+} - \gamma q_{j+}) \leq \delta_i \leq \max_j(a_{jj} + \sum_{k\neq j}|a_{jk}|) = \max_j(q_{j+} + \gamma q_{j+})$, where q_{j+} is as defined in (4). In addition, under adjacent neighborhood and regular grid, one site has at least 2 neighbors and at most 4 neighbors, indicating $2 \leq q_{j+} \leq 4$. Therefore, $2(1-\gamma) \leq \delta_i \leq 4(1+\gamma)$ for $i = 1, \ldots, n$. As the eigenvalues of $\mathbf{R} = \mathbf{A}^{-1}$ are δ_i^{-1}, the eigenvalues of $\mathbf{\Lambda} = \theta\mathbf{R} + \sigma_\epsilon^2\mathbf{I}$ are hence $\theta\delta_i^{-1} + \sigma_\epsilon^2$. Therefore, for the CAR model,

$$\lambda_n \leq \frac{\theta}{2(1-\gamma)} + \sigma_\epsilon^2, \tag{S1.4}$$

where γ is a constant in $(-1, 1)$.

In view of (S1.1), (S1.2), (S1.3) and (S1.4), λ_n is bounded (in a compact parameter set), leading to $\limsup \lambda_n < \infty$. Now consider the spectrum of $\partial\mathbf{\Lambda}/\partial\theta$, denoted by $|\lambda_n^1|$ As $\mathbf{\Lambda}_1 = \partial\mathbf{\Lambda}/\partial\theta = \mathbf{R}$, it follows immediately $\limsup |\lambda_n^1| < \infty$. Also, $\mathbf{\Lambda}_2 = \partial\mathbf{\Lambda}/\partial\sigma_\epsilon = 2\sigma_\epsilon\mathbf{I}$, whose spectrum is $2\sigma_\epsilon$, a finite constant. In addition, $\partial^2\mathbf{\Lambda}/\partial\theta^2 = \partial^2\mathbf{\Lambda}/\partial\theta\partial\sigma_\epsilon = 0$ and $\partial^2\mathbf{\Lambda}/\partial\sigma_\epsilon^2 = 2\mathbf{I}$, whose spectra trivially satisfy (c.1).

We now verify (c.2). As the diagonal elements of correlation matrix $\mathbf{R}$ are 1's, $||\mathbf{\Lambda}_1|| = ||\mathbf{R}|| \geq \sqrt{n}$. In addition, $||\mathbf{\Lambda}_2|| = 2\sigma_\epsilon\sqrt{n}$. Hence, (c.2) is satisfied with $\delta = 1/2$.

S2. Proof of Asymptotic Bias of the Naive Regression Coefficients When Measurement Error is Ignored (Theorem 1)

(i) Under conditions (c.1)-(c.4), Lemma 4 of Sweeting (1980) implies that a maximizer to (10) or a solution to (11) exists. Furthermore, Theorem 2 of Sweeting (1980) implies such a solution converges in probability to the asymptotic solution to (12). Now let the probability limits of the naive estimators for β_0, β_x, σ_ϵ, and θ (as in $\mathbf{V}(\theta)$) be $\beta_{0,\text{naive}}$, $\beta_{x,\text{naive}}$, $\sigma_{\epsilon,\text{naive}}$, and θ_{naive} respectively.

Then they should satisfy the following probability limit of score equations for regression coefficients.

$$\lim \frac{1}{n} E((\mathbf{1}\,\mathbf{W})^T(\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_{\epsilon,\text{naive}}\mathbf{I})^{-1}(\mathbf{Y} - (\mathbf{1}\,\mathbf{W})(\beta_{0,\text{naive}}, \beta_{x,\text{naive}})^T)) = 0 \tag{S2.1}$$

Using the equality

$$E(\boldsymbol{\psi}^T \boldsymbol{B} \boldsymbol{\psi}) = \text{tr}(\boldsymbol{B}\text{cov}(\boldsymbol{\psi})) + E(\boldsymbol{\psi})^T \boldsymbol{B} E(\boldsymbol{\psi})$$

for any random vector $\boldsymbol{\psi}$ (which can be $\mathbf{X}$ or $\mathbf{W}$ in this case), we have

$$\beta_0 + \beta_x \alpha_0 = \beta_{0,\text{naive}} + \beta_{x,\text{naive}}\alpha_0 \tag{S2.2}$$

$$\begin{aligned}
&\beta_x \lim\{\frac{1}{n}(\text{tr}((\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_\epsilon \mathbf{I})^{-1}(\mathbf{\Sigma}(\boldsymbol{\zeta}) + \sigma^2_e \mathbf{I})) + \alpha_0^2 \mathbf{1}^T(\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_\epsilon \mathbf{I})^{-1}\mathbf{1})\} = \\
&\quad \beta_{0,\text{naive}}\alpha_0 \lim \frac{1}{n}\mathbf{1}^T(\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_\epsilon \mathbf{I})^{-1}\mathbf{1} + \beta_{x,\text{naive}} \lim\, \{\frac{1}{n}(\text{tr}((\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_\epsilon \mathbf{I})^{-1} \\
&\quad (\mathbf{\Sigma}(\boldsymbol{\zeta}) + (\sigma^2_e + \sigma^2_U)\mathbf{I})) + \alpha_0^2 \mathbf{1}^T(\mathbf{V}(\theta_{\text{naive}}) + \sigma^2_\epsilon \mathbf{I})^{-1}\mathbf{1})\}.
\end{aligned} \tag{S2.3}$$

Solving the above equations for $\beta_{0,\text{naive}}$, $\beta_{x,\text{naive}}$, we have the results. $\beta_{x,\text{naive}}$, we have the results.

(ii) Use the definition of λ_* and the fact that all the matrices involved in λ_* can be diagonalized by the same orthogonal matrix.

(iii) First consider the CAR correlation matrix. Under adjacent neighborhood and regular (square) grid, one site has at most 4 neighbors. Using the same argument right above (S1.4) in section A.0, we have that $0 < \delta_l < 4(1+\gamma)$. As the right hand side of (16) is a non-increasing function for each δ_l (with a fixed n), appplying this inequality to (16) yields (17).

For the exponential decaying correlation matrix, denote the eigenvalues of the correlation matrix $\mathbf{R}$ by $\tilde{\delta}_l, l = 1, \ldots, n$. Then $\tilde{\delta}_l$ can not exceed the row sum norm (Theorem 5.6.7 of Graybill, 1969). Using (S1.1), $\tilde{\delta}_l < 4(\frac{1-e^{-\sqrt{n/2}}}{1-e^{-1/\sqrt{2}}})^2 < \frac{4}{(1-e^{-1/\sqrt{2}})^2}$. Hence, $\delta_l = \tilde{\delta}_l^{-1} > \frac{1}{4}(1-e^{-1/\sqrt{2}})^2$. Using this inequality in (16) leads to the result of (18). Similarly, with the “Gaussian” spatial correlation, noting (S1.2) and (S1.3) leads to the same bound in (18).

S3. Proof of the Asymptotic Bias of the Naive Variance Components when Measurement Error is ignored (Theorem 2)

Let the probability limits of the naive estimators be $\beta_{0,\text{naive}}$, $\beta_{x,\text{naive}}$, and $\boldsymbol{\vartheta}_{\text{naive}} = (\theta_{\text{naive}}, \sigma^2_{\epsilon,\text{naive}}) \overset{def}{=} (\vartheta_1, \vartheta_2)$. Then they are solutions of (S2.2) and (S2.3), and

$$\lim \frac{1}{2n}\{E((\mathbf{Y} - \beta_{0,\text{naive}}\mathbf{1} - \beta_{x,\text{naive}}\mathbf{W})^T \mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\vartheta_j}\mathbf{S}^{-1}$$
$$(\mathbf{Y} - \beta_{0,\text{naive}}\mathbf{1} - \beta_{x,\text{naive}}\mathbf{W})) - \text{tr}(\mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\vartheta_j})\} = 0, \qquad \text{(S3.1)}$$

where $\mathbf{S} = \theta_{\text{naive}}\mathbf{R} + \sigma^2_{\epsilon,\text{naive}}\mathbf{I}$ and $j = 1, 2$.

Let $\mathbf{T} = \mathbf{Y} - \beta_{0,\text{naive}}\mathbf{1} - \beta_{x,\text{naive}}\mathbf{W}$. Using (S2.1) yields $E(\mathbf{T}) = 0$. In addition, using (S3.1) and the fact that $\frac{\partial \mathbf{S}}{\partial \theta_{\text{naive}}} = \mathbf{R}$, $\frac{\partial \mathbf{S}}{\partial \sigma^2_{\epsilon,\text{naive}}} = \mathbf{I}$, we obtain that

$$\lim \frac{1}{n}\text{tr}(\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}\text{cov}(\mathbf{T})) = \lim \frac{1}{n}\text{tr}(\mathbf{S}^{-1}\mathbf{R}),\ \lim \frac{1}{n}\text{tr}(\mathbf{S}^{-2}\text{cov}(\mathbf{T})) = \lim \frac{1}{n}\text{tr}(\mathbf{S}). \qquad \text{(S3.2)}$$

However $\text{cov}(\mathbf{T})$ is of the same form as $\mathbf{S}$, i.e., a linear combination of $\mathbf{R}$ and $\mathbf{I}$, since

$$\begin{aligned}
\text{cov}(\mathbf{T}) &= \text{cov}(\mathbf{Y} - \beta_{x,\text{naive}}\mathbf{W}) \\
&= \text{cov}(\mathbf{Y}) + \beta^2_{x,\text{naive}}\text{cov}(\mathbf{W}) - 2\beta_{x,\text{naive}}\text{cov}(\mathbf{Y}, \mathbf{W}) \\
&= (\sigma^2_{\Sigma}(\beta_x - \beta_{x,\text{naive}})^2 + \theta)\mathbf{R} + (\sigma^2_e(\beta_x - \beta_{x,\text{naive}})^2 + \sigma^2_\epsilon + \beta^2_{x,\text{naive}}\sigma^2_U)\mathbf{I}.
\end{aligned}$$

Therefore for (S3.2) to hold, $\text{cov}(\mathbf{T}) = \mathbf{S}$. Compare the coefficients of $\mathbf{R}$ and $\mathbf{I}$, (19) follows.

S4. Proof of the Consistency and Asymptotic Normality of the MLEs (Theorem 3)

The proof centers on verifying the sufficient conditions, along the line of Mardia and Marshall (1984), that allow the use of Sweeting (1980) concerning consistency and asymptotic normality of MLEs for Gaussian models, as $\mathbf{Y}$ and $\mathbf{W}$ jointly follow a multivariate normal distribution (7). However, the variance-covariance matrix of the observed $(\mathbf{Y}, \mathbf{W})$, denoted by $\boldsymbol{\Lambda}$, involves regression coefficients. Hence, the regression coefficients and the variance components are not orthogonal. It is thus difficult to directly apply Mardia and Marshall's (1984) results, which required such orthogonality to ensure that the information matrix is block diagonal. To circumvent this problem, we carry out the following reparameterization.

$$\begin{aligned} \alpha_0^* = \alpha_0, \quad \boldsymbol{\alpha}_z^* = \boldsymbol{\alpha}_z, \quad \beta_0^* = \beta_0 + \beta_x \alpha_0, \\ \boldsymbol{\beta}_z^* = \beta_x \boldsymbol{\alpha}_z + \boldsymbol{\beta}_z, \quad \sigma_1^* = \beta_x, \quad \boldsymbol{\theta}^* = \boldsymbol{\theta}, \\ \boldsymbol{\zeta}^* = \boldsymbol{\zeta}, \quad \sigma_e^* = \sigma_e, \quad \sigma_\epsilon^* = \sigma_\epsilon. \end{aligned} \tag{S4.1}$$

In the ensuing development, we use $\boldsymbol{\Omega} = (\alpha_0, \boldsymbol{\alpha}_z, \beta_0, \beta_x, \boldsymbol{\beta}_z, \boldsymbol{\theta}, \boldsymbol{\zeta}, \sigma_e, \sigma_\epsilon)$ to denote the collection of original parameters, and $\boldsymbol{\Omega}^* = (\alpha_0^*, \boldsymbol{\alpha}_z^*, \beta_0^*, \beta_z^*, \sigma_1^*, \boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \sigma_e^*, \sigma_\epsilon^*)$ to denote the collection of new parameters. Under such a reparameterization, the joint distribution of $(\mathbf{Y}, \mathbf{W}|\mathbf{Z})$ is specified by $\boldsymbol{\Omega}^*$, with likelihood

$$\ell(\mathbf{Y}, \mathbf{W}|\mathbf{Z}) \quad = \quad -\frac{(2n)}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Lambda}| - \frac{1}{2}\begin{pmatrix} \mathbf{Y} - \mu_y \\ \mathbf{W} - \mu_w \end{pmatrix}^T \boldsymbol{\Lambda}^{-1} \begin{pmatrix} \mathbf{Y} - \mu_y \\ \mathbf{W} - \mu_w \end{pmatrix} \tag{S4.2}$$

where $\mu_y = \beta_0^* \mathbf{1} + \mathbf{Z}\boldsymbol{\beta}_z^*$, $\mu_w = \alpha_0^* \mathbf{1} + \mathbf{Z}\boldsymbol{\alpha}_z^*$ and

$$\begin{aligned} \boldsymbol{\Lambda} &= \operatorname{cov}(\mathbf{Y}, \mathbf{W}|\mathbf{Z}) \\ &= \begin{pmatrix} (\sigma_1^*)^2 \boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + \mathbf{V}(\boldsymbol{\theta}^*) + \{(\sigma_1^*)^2(\sigma_e^*)^2 + (\sigma_\epsilon^*)^2\}\mathbf{I} & \sigma_1^*\{\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I}\} \\ \sigma_1^*\{\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I}\} & \boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + \{(\sigma_e^*)^2 + \sigma_U^2\}\mathbf{I} \end{pmatrix}. \end{aligned}$$

Denote by $\boldsymbol{\beta}_* = (\alpha_0^*, \boldsymbol{\alpha}_z^*, \beta_0^*, \boldsymbol{\beta}_z^*)$ the vector of (new) regression coefficients and $\boldsymbol{\theta}_* = (\sigma_1^*, \boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \sigma_e^*, \sigma_\epsilon^*)$ the vector of (new) variance components. Direct computation yields $E(-\partial^2 \ell / \partial \boldsymbol{\beta}_* \partial \boldsymbol{\theta}_*^T) = 0$. Hence, the reparameterization leads to orthogonality of the regression coefficients $\boldsymbol{\beta}_*$ and the variance components $\boldsymbol{\theta}_*$, which fits the analytical framework of Mardia and Marshall (1984). Hence, in order to show consistency and asymptotic normality of the maximum likelihood estimator, it suffices to show that the following modified regularity conditions of Mardia and Marshall (1984) hold [which are similar to conditions (c.1)-(c.4) considered for the naive estimator in Section 3].

For notational ease, we denote $\boldsymbol{\theta}_*$ as $\boldsymbol{\theta}_* = (\vartheta_1^*, \ldots, \vartheta_q^*) \equiv (\sigma_1^*, \boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \sigma_e^*, \sigma_\epsilon^*)$. Denote by $\boldsymbol{\Lambda}_i^* = \partial / \partial \vartheta_i^* \boldsymbol{\Lambda}(\boldsymbol{\theta}_*)$ and $\boldsymbol{\Lambda}_{ij}^* = \partial^2 / \partial \vartheta_i^* \partial \vartheta_j^* \boldsymbol{\Lambda}(\boldsymbol{\theta}_*)$, where the differentiation is element-wise. Now let $\lambda_1 \leq \ldots \leq \lambda_n$ be the eigen-values of $\boldsymbol{\Lambda}$ and let those of $\mathbf{\Lambda}_i^*$ and $\mathbf{\Lambda}_{ij}^*$ be λ_k^i and λ_k^{ij} for $k = 1, \ldots, n$ respectively, with $|\lambda_1^i| \leq \ldots \leq |\lambda_n^i|$ and $|\lambda_1^{ij}| \leq \ldots \leq |\lambda_n^{ij}|$ for $i, j = 1, \ldots, q$. The sufficient conditions are as follows.

(h.1) $\limsup \lambda_n < \infty, \limsup |\lambda_n^i| < \infty, \limsup |\lambda_n^{ij}| < \infty$, for all $i, j = 1, \ldots q$.

(h.2) $||\mathbf{\Lambda}_i^*||^{-2} = O(n^{-\frac{1}{2}-\delta})$ for some $\delta > 0$ for $i = 1, \ldots q$.

(h.3) $\mathbf{A} = (a_{ij})$ is invertible, where for all ij, $a_{ij} = \{t_{ij}/(t_{ii}t_{jj})^{1/2}\}$ exists and $t_{ij} = tr(\mathbf{\Lambda}^{-1}\mathbf{\Lambda}_i^*\mathbf{\Lambda}^{-1}\mathbf{\Lambda}_j^*)$.

(h.4) $\lim(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} = 0$ in probability.

To show (h.1) hold, we again use the matrix norm property that the spectral radius of any matrix is bounded by its row sum norm. Therefore,

$$\begin{aligned}\lambda_n \leq\; & (\sigma_1^*)^2||\mathbf{\Sigma}(\boldsymbol{\zeta}^*)||_\infty + ||\mathbf{V}(\boldsymbol{\theta}^*)||_\infty + \{(\sigma_1^*)^2(\sigma_e^*)^2 + (\sigma_\epsilon^*)^2\} \\ & + \sigma_1^*\{||\mathbf{\Sigma}(\boldsymbol{\zeta}^*)||_\infty + (\sigma_e^*)^2\} + ||\mathbf{\Sigma}(\boldsymbol{\zeta}^*)||_\infty + (\sigma_e^*)^2 + \sigma_U^2.\end{aligned}$$

As the row sum norms of $\mathbf{\Sigma}(\boldsymbol{\zeta}^*)$ and $\mathbf{V}(\boldsymbol{\theta}^*)$ are finite under the CAR model, the exponential model, the Gaussian model as shown in (S1.4), (S1.1),(S1.2) and (S1.3), along with the assumption of (d.1), λ_n is bounded (when n is sufficiently large), leading to $\limsup \lambda_n < \infty$. Now consider the spectrum of $\partial\mathbf{\Lambda}/\partial\sigma_1^*$, denoted by $|\lambda_n^1|$. Indeed,

$$|\lambda_n^1| \leq 2\sigma_1^*|\mathbf{\Sigma}(\boldsymbol{\zeta}^*)||_\infty + 2\sigma_1^*(\sigma_e^*)^2 + ||\mathbf{\Sigma}(\boldsymbol{\zeta}^*)||_\infty + (\sigma_e^*)^2,$$

leading to $\limsup |\lambda_n^1| < \infty$. Similarly, we can show $\limsup |\lambda_n^i| < \infty, \limsup |\lambda_n^{ij}| < \infty$, for all $i, j = 1, \ldots q$. Hence condition (h.1) is verified.

We now verify (h.2). First consider

$$\begin{aligned}\mathbf{\Lambda}_1^* = \frac{\partial}{\partial\sigma_1^*}\mathbf{\Lambda} &= \begin{pmatrix} 2\sigma_1^*\mathbf{\Sigma}(\boldsymbol{\zeta}^*) + 2\sigma_1^*(\sigma_e^*)^2\mathbf{I} & \mathbf{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I} \\ \mathbf{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2\sigma_1^*\mathbf{\Sigma}(\boldsymbol{\zeta}^*) & \mathbf{\Sigma}(\boldsymbol{\zeta}^*) \\ \mathbf{\Sigma}(\boldsymbol{\zeta}^*) & 0 \end{pmatrix} + \begin{pmatrix} 2\sigma_1^*(\sigma_e^*)^2\mathbf{I} & (\sigma_e^*)^2\mathbf{I} \\ (\sigma_e^*)^2\mathbf{I} & 0 \end{pmatrix}. \quad \text{(S4.3)}\end{aligned}$$

We denote the first matrix in (S4.3) by $\mathbf{\Lambda}_{1,1}^*$ and the second by $\mathbf{\Lambda}_{1,2}^*$. Some algebra yields that $||\mathbf{\Lambda}_{1,2}^*||^2 = \{4(\sigma_1^*)^2(\sigma_e^*)^4 + 2(\sigma_e^*)^4\}n$. Using the definition of the matrix norm $||\cdot||$, we have that

$$||\mathbf{\Lambda}_1^*||^2 = ||2\sigma_1^*\mathbf{\Sigma}(\boldsymbol{\zeta}^*) + 2\sigma_1^*(\sigma_e^*)^2\mathbf{I}||^2 + 2||\mathbf{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I}||^2.$$

Since the diagonal elements of $\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*)$ and $\mathbf{I}$ are all nonnegative (and hence the diagonal elements of $\sigma_1^*\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*)$ and $\sigma_1^*(\sigma_e^*)^2\mathbf{I}$ have the same sign), and the off-diagonal elements of $\mathbf{I}$ are zero, it follows that $||2\sigma_1^*\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + 2\sigma_1^*(\sigma_e^*)^2\mathbf{I}||^2 \geq ||2\sigma_1^*(\sigma_e^*)^2\mathbf{I}||^2$ and $||\boldsymbol{\Sigma}(\boldsymbol{\zeta}^*) + (\sigma_e^*)^2\mathbf{I}||^2 \geq ||(\sigma_e^*)^2\mathbf{I}||^2$. Hence,

$$||\boldsymbol{\Lambda}_1^*||^2 \geq ||2\sigma_1^*(\sigma_e^*)^2\mathbf{I}||^2 + 2||(\sigma_e^*)^2\mathbf{I}||^2 = ||\boldsymbol{\Lambda}_{1,2}^*||^2.$$

That is, we have obtained that $||\boldsymbol{\Lambda}_1^*||^2 \geq \{4(\sigma_1^*)^2(\sigma_e^*)^4 + 2(\sigma_e^*)^4\}n$, or $||\boldsymbol{\Lambda}_1^*||^{-2} \leq \mathcal{C}\times n^{-1}$, where $\mathcal{C}$ is a positive constant (not depending on n). Then it follows that $||\boldsymbol{\Lambda}_1^*||^{-2} = O(n^{-1})$. Taking derivatives of $\boldsymbol{\Lambda}$ with respect to the other variance components, and using the same arguments [i.e. the similar matrix decomposition as in (S4.3)], we will have that $||\boldsymbol{\Lambda}_i^*||^{-2} = O(n^{-1})$ for $i = 1, \ldots, q$. Hence, condition (h.2) holds with $\delta = 1/2$.

We are in a position to verify (h.3). First, define $\mathcal{T} = -E(\partial^2\ell/\partial\boldsymbol{\theta}_*\partial\boldsymbol{\theta}_*^T)$, where the derivative and the expectation are performed under the true parameters. By the definition of t_{ij}, it follows that $\mathcal{T} = (t_{ij})_{q\times q}$, where q is the dimension of the variance components $\boldsymbol{\theta}_*$. Moreover,

$$\mathcal{T} = E\left\{\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}_*}\right)\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}_*}\right)^T\right\} = var\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}_*}\right).$$

As the variance components of $\boldsymbol{\theta}_*$ are not linearly dependent, $\mathcal{T}$ is positive definite. Hence $t_{ii} > 0$ and hence $a_{ij} = t_{ij}/(t_{ii}t_{jj})^{1/2}$ is well defined. Furthermore, $\mathbf{A} \overset{def}{=} (a_{ij})_{q\times q} = \mathbf{D}^{1/2}\mathbf{T}\mathbf{D}^{1/2}$, where $\mathbf{D} = diag(t_{11}, \ldots, t_{qq})$. Hence, $\mathbf{A} = (a_{ij})$ is invertible, which verifies (h.3).

Finally, (h.4) follows immediately as $\lim(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} = \lim \mathcal{Z}_0^{-1}/n = 0$, based on the regularity condition (d.2) for the observed covariates.

With the sufficient conditions (h.1)-(h.4) checked, the MLE (denoted by $\widehat{\boldsymbol{\Omega}^*}$) for the transformed parameter $\boldsymbol{\Omega}^* = (\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ are consistent and asymptotically normal. That is, $\widehat{\boldsymbol{\Omega}^*} - \boldsymbol{\Omega}_0^* \sim N(0, \boldsymbol{\Gamma}_*^{-1})$, where $\sim$ corresponds to asymptotic equivalence in distribution,
$\boldsymbol{\Gamma}_* = E_{\boldsymbol{\Omega}_0^*}\{-\partial^2\ell/\partial\boldsymbol{\Omega}^*\partial(\boldsymbol{\Omega}^*)^T\}$ and $\boldsymbol{\Omega}_0^*$ is the truth under reparameterization (Mardia and Marshall, 1984). Here, ℓ is as defined in (S4.2) or, equally, in (7). Obviously, the reparameterization from the original parameter $\boldsymbol{\Omega}$ to $\boldsymbol{\Omega}^*$ in (S4.1) is continuously invertible and differentiable. That is, $\boldsymbol{\Omega} = \mathcal{F}(\boldsymbol{\Omega}^*)$ for a one-one

and differentiable function $\mathcal{F}(\cdot)$. Indeed, the components of $\mathcal{F}(\cdot)$ is as follows.

$$\alpha_0 = \alpha_0^*, \quad \boldsymbol{\alpha}_z = \boldsymbol{\alpha}_z^*, \quad \beta_0 = \beta_0^* - \sigma_1^*\alpha_0^*,$$
$$\boldsymbol{\beta}_z = \boldsymbol{\beta}_z^* - \sigma_1^*\boldsymbol{\alpha}_z^*, \quad \beta_x = \sigma_1^*, \quad \boldsymbol{\theta} = \boldsymbol{\theta}^*,$$
$$\boldsymbol{\zeta} = \boldsymbol{\zeta}^*, \quad \sigma_e = \sigma_e^*, \quad \sigma_\epsilon = \sigma_\epsilon^*.$$

By the reparameterization-invariance principle of the maximum likelihood estimator (e.g. Lehman and Casella, 1998), $\widehat{\boldsymbol{\Omega}} \overset{def}{=} \mathcal{F}(\widehat{\boldsymbol{\Omega}^*})$ is the MLE of the original parameter $\boldsymbol{\Omega}$. Further, as $\mathcal{F}(\cdot)$ is smooth, $\widehat{\boldsymbol{\Omega}}$ is consistent and asymptotically normal. Using the delta method, the variance of $\mathcal{F}(\widehat{\boldsymbol{\Omega}^*})$ is approximately equal to $\{\frac{\partial \mathcal{F}}{\partial \boldsymbol{\Omega}^*}\}^T \boldsymbol{\Gamma}_*^{-1}\{\frac{\partial \mathcal{F}}{\partial \boldsymbol{\Omega}^*}\} = \{E_{\boldsymbol{\Omega}_0}(-\partial^2 \ell/\partial\boldsymbol{\Omega}\partial\boldsymbol{\Omega}^T)\}^{-1}$, where the last equality is due to the chain rule (see Schervish, 1995). Indeed, $\boldsymbol{\Gamma} \overset{def}{=} E_{\boldsymbol{\Omega}_0}\{-\partial^2\ell/\partial\boldsymbol{\Omega}\partial\boldsymbol{\Omega}^T\}$ is the information under the original parameter $\boldsymbol{\Omega}$. Therefore, $\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0 \sim N(0, \boldsymbol{\Gamma}^{-1})$ or equivalently $\boldsymbol{\Gamma}^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0) \to N(0, \mathbf{I}_p)$ in distribution, where $\boldsymbol{\Gamma}^{1/2}$ is the Cholesky decomposition of $\boldsymbol{\Gamma}$ and $\mathbf{I}_p$ is the identity matrix of dimension of p, the dimension of $\boldsymbol{\Omega}_0$.

S5. Implementation of the EM algorithm

The E step detailed in Section 4 needs the following the expectations of quantities conditional on the observed data ($\mathbf{Y}$, $\mathbf{W}$, $\mathbf{Z}$) and current values of the parameter estimates.

$$E(\mathbf{1}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = \mathbf{1}^T E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$E(\mathbf{X}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})^T E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + \mathrm{tr}(\mathrm{cov}(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}))$$
$$E(\mathbf{Z}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = \mathbf{Z}^T E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$E(\mathbf{X}^T(\mathbf{Y}-\mathbf{b})|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})^T\mathbf{Y} - E(\mathbf{X}^T\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$E(\mathbf{b}^T\mathbf{V}^{-1}\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})^T\mathbf{V}^{-1}E(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + \mathrm{tr}(\mathrm{cov}(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})\mathbf{V}^{-1})$$
$$E(\|\mathbf{Y} - \hat{\beta}_x^{(t+1)}\mathbf{X} - \mathbf{Z}\beta_z^{(t+1)} - \mathbf{b}\|^2 \mid \mathbf{Y},\mathbf{W},\hat{\boldsymbol{\theta}}^{(t)}) = (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_z^{(t+1)})^T(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_z^{(t+1)})$$
$$-2(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_z^{(t+1)})^T(E(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + \hat{\beta}_x^{(t+1)}E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})) + E(\mathbf{b}^T\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$+(\hat{\beta}_x^{(t+1)})^2E(\mathbf{X}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + 2\hat{\beta}_x^{(t+1)}E(\mathbf{b}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$E((\mathbf{X}-\mathbf{a})^T(\mathbf{X}-\mathbf{a})|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E(\mathbf{X}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + E(\mathbf{a}^T\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$-2E(\mathbf{a}^T\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$$
$$E(\mathbf{a}^T\mathbf{V}^{-1}\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})^T\mathbf{V}^{-1}E(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) +$$

$$\mathrm{tr}(\mathrm{cov}(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})\mathbf{V}^{-1})$$

$$E(\|\mathbf{X}-\hat{\alpha}_0^{(t+1)}\mathbf{1}-\mathbf{Z}\boldsymbol{\alpha}_z^{(t+1)}-\mathbf{a}\|^2|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) = E((\mathbf{X}-\mathbf{a})^T(\mathbf{X}-\mathbf{a})|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) + n\hat{\alpha}_0^{(t+1)}$$
$$-2\hat{\alpha}_0^{(t+1)}(E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) - E(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})) + (\boldsymbol{\alpha}_z^{(t+1)})^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\alpha}_z^{(t+1)}$$
$$+2\hat{\alpha}_0^{(t+1)}\mathbf{1}^T\mathbf{Z}\boldsymbol{\alpha}_z^{(t+1)} - 2(\hat{\boldsymbol{\alpha}}_z^{(t+1)})^T\mathbf{Z}^T(E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}) - E(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})),$$

where $E(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$,$E(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$, $E(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$, $\mathrm{cov}(\mathbf{a}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$, $\mathrm{cov}(\mathbf{b}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$, $\mathrm{cov}(\mathbf{X}|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)})$ can be obtained from

$$\mathrm{cov}\left(\begin{array}{c}\mathbf{X}\\ \mathbf{a}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) = \begin{pmatrix} (\beta_x^{(t)})^2(\theta^{(t)}\mathbf{R}+\sigma_\epsilon^{2(t)}\mathbf{I})^{-1} + (\frac{1}{\sigma_U^2}+\frac{1}{\sigma_e^{2(t)}})\mathbf{I} & -\frac{1}{\sigma_e^{2(t)}}\mathbf{I} \\ -\frac{1}{\sigma_e^{2(t)}}\mathbf{I} & (\sigma_\Sigma^{2(t)}\mathbf{V})^{-1} + \frac{1}{\sigma_e^{2(t)}}\mathbf{I} \end{pmatrix}^{-1}$$

$$E\left(\begin{array}{c}\mathbf{X}\\ \mathbf{a}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) = \mathrm{cov}\left(\begin{array}{c}\mathbf{X}\\ \mathbf{a}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) \times \begin{pmatrix} \beta_x^{(t)}(\theta(t)\mathbf{V}+\sigma_\epsilon^{2(t)}\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}_z^{(t)}) + \frac{\mathbf{W}}{\sigma_U^2} + \frac{\alpha_0^{(t)}}{\sigma_e^{2(t)}}\mathbf{1} \\ -\frac{1}{\sigma_e^{2(t)}}(\alpha_0^{(t)}\mathbf{1}+\mathbf{Z}\boldsymbol{\alpha}_z^{(t)}) \end{pmatrix}$$

$$\mathrm{cov}\left(\begin{array}{c}\mathbf{X}\\ \mathbf{b}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) = \begin{pmatrix} (\sigma_\Sigma^{2(t)}\mathbf{V}+\sigma_e^{2(t)}\mathbf{I})^{-1} + (\frac{1}{\sigma_U^2}+\frac{(\beta_x^{(t)})^2}{\sigma_\epsilon^{2(t)}})\mathbf{I} & \frac{\beta_x}{\sigma_e^{2(t)}}\mathbf{I} \\ \frac{\beta_x}{\sigma_e^{2(t)}}\mathbf{I} & (\theta^{(t)}\mathbf{V})^{-1} + \frac{1}{\sigma_\epsilon^{2(t)}}\mathbf{I} \end{pmatrix}^{-1}$$

$$E\left(\begin{array}{c}\mathbf{X}\\ \mathbf{b}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) = \mathrm{cov}\left(\begin{array}{c}\mathbf{X}\\ \mathbf{b}\end{array}\middle|\mathbf{Y},\mathbf{W},\mathbf{Z},\hat{\boldsymbol{\theta}}^{(t)}\right) \times \begin{pmatrix} \frac{\beta_x}{\sigma_\epsilon^{2(t)}}(\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}_z) + \frac{\mathbf{W}}{\sigma_U^2} + (\sigma_\Sigma^{2(t)}\mathbf{V}+\sigma_e^{2(t)}\mathbf{I})^{-1}(\alpha_0^{(t)}\mathbf{1}+\mathbf{Z}\boldsymbol{\alpha}_z^{(t)}) \\ \frac{1}{\sigma_\epsilon}(\mathbf{Y}-\mathbf{Z}\boldsymbol{\beta}_z) \end{pmatrix}.$$

Hence, the E steps can be easily implemented since all the quantities involved have closed-form and no numerical integrations are needed.