Supplementary Materials for "MM Algorithms for Variance Components Models"

Liuyi Hu Hua Zhou Jin Zhou Kenneth Lange

S.1 Proof for Lemma 2

Proof. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random normal vector with mean $\mathbf{0}$ and positive definite covariance matrix $\boldsymbol{A}$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be a random normal vector independent of $\boldsymbol{X}$ with mean $\mathbf{0}$ and positive semidefinite covariance matrix $\boldsymbol{B}$ having positive diagonal entries. Then $\boldsymbol{Z} = \boldsymbol{X} \odot \boldsymbol{Y}$ has covariances $\mathrm{E}(Z_i Z_j) = \mathrm{E}(X_i Y_i X_j Y_j) = \mathrm{E}(X_i X_j)\mathrm{E}(Y_i Y_j) = a_{ij} b_{ij}$. It follows that $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{A} \odot \boldsymbol{B}$. To show $\boldsymbol{A} \odot \boldsymbol{B}$ is positive definite, suppose on the contrary that $\boldsymbol{v}^T(\boldsymbol{A} \odot \boldsymbol{B})\boldsymbol{v} = \mathrm{Var}(\boldsymbol{v}^T \boldsymbol{Z}) = 0$ for some $\boldsymbol{v} \neq \mathbf{0}$. Then

$$0 \;=\; \mathrm{Var}(\boldsymbol{v}^T \boldsymbol{Z}) = \mathrm{E}\Big(\sum_i v_i X_i Y_i \Big)^2 = \mathrm{E}\Big[\Big(\sum_i v_i X_i Y_i \Big)^2 \mid \boldsymbol{Y} \Big] = \mathrm{E}[(\boldsymbol{v} \odot \boldsymbol{Y})^T \boldsymbol{A} (\boldsymbol{v} \odot \boldsymbol{Y})]$$

implies $\boldsymbol{v} \odot \boldsymbol{Y} = \mathbf{0}$ with probability 1. Since $\boldsymbol{v} \neq \mathbf{0}$, $Y_i = 0$ with probability 1 for some i. This contradicts the assumption $b_{ii} = \mathrm{Var}(Y_i) > 0$ for all i. □

S.2 Objective Values in ANOVA Simulation Example

Table 1 summarizes the converged objective values for the two-way ANOVA example in Section 3. Reported in table are average and standard error based on 50 simulation replicates.

S.3 Proof of Theorem 1

We need three technical Lemmas to show the global convergence result in Theorem 1.

Lemma S.1. *Under Assumption 1 or 2, the log-likelihood function (1) is coercive in the sense that the super-level set $S_c = \{\boldsymbol{\sigma}^2 \geq \mathbf{0} : L(\boldsymbol{\sigma}^2) \geq c\}$ is compact for every c.*

Proof. Let us first prove the assertion when all of the covariance matrices $\boldsymbol{V}_i$ are positive definite. If we set $r = \|\boldsymbol{\sigma}^2\|_1$ and $\alpha_i = r^{-1}\sigma_i^2$ for each i, then the log-likelihood satisfies

$$L(\boldsymbol{\sigma}^2) \;=\; -\frac{n}{2} \ln r - \frac{1}{2} \ln \det \Big(\sum_{i=1}^m \alpha_i \boldsymbol{V}_i \Big) - \frac{1}{2r} \boldsymbol{y}^T \Big(\sum_{i=1}^m \alpha_i \boldsymbol{V}_i \Big)^{-1} \boldsymbol{y}.$$

The functions $\ln \det \Big(\sum_{i=1}^m \alpha_i \boldsymbol{V}_i \Big)$ and $\boldsymbol{y}^T \Big(\sum_{i=1}^m \alpha_i \boldsymbol{V}_i \Big)^{-1} \boldsymbol{y}$ of $\boldsymbol{\alpha}$ are defined and continuous on the unit simplex and hence bounded there. The dominant term $-\frac{n}{2} \ln r$ of the loglikelihood tends to $-\infty$ as r tends to ∞.

Table 1: MM, EM, Fisher scoring, and `lme4` converge to similar objective values. Shown below are average objective values for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

σ_1^2/σ_e^2	Method	c = # observations per combination			
		5	10	20	50
0.00	MM	-176.67(7.94)	-353.59(10.19)	-713.42(14.90)	-1776.40(25.02)
	EM	-176.68(7.94)	-353.60(10.18)	-713.43(14.90)	-1776.41(25.02)
	FS	-176.67(7.94)	-353.59(10.19)	-713.42(14.90)	-1776.40(25.02)
	lme4	-176.67(7.94)	-353.59(10.19)	-713.42(14.90)	-1776.40(25.02)
0.05	MM	-181.06(7.24)	-365.39(10.92)	-722.16(15.36)	-1794.26(25.96)
	EM	-181.06(7.24)	-365.39(10.92)	-722.16(15.36)	-1794.26(25.96)
	FS	-181.06(7.24)	-365.39(10.92)	-722.16(15.36)	-1794.26(25.96)
	lme4	-181.06(7.24)	-365.39(10.92)	-722.16(15.36)	-1794.26(25.96)
0.10	MM	-185.10(6.77)	-368.83(10.28)	-726.21(13.35)	-1813.88(20.24)
	EM	-185.10(6.77)	-368.83(10.28)	-726.21(13.35)	-1813.88(20.24)
	FS	-185.10(6.77)	-368.83(10.28)	-726.21(13.35)	-1813.88(20.24)
	lme4	-185.10(6.77)	-368.83(10.28)	-726.21(13.35)	-1813.88(20.24)
1.00	MM	-204.05(8.18)	-392.35(10.94)	-754.95(13.93)	-1831.11(22.63)
	EM	-204.05(8.18)	-392.35(10.94)	-754.95(13.93)	-1831.12(22.63)
	FS	-204.05(8.18)	-392.35(10.94)	-754.95(13.93)	-1831.11(22.63)
	lme4	-204.05(8.18)	-392.35(10.94)	-754.95(13.93)	-1831.11(22.63)
10.00	MM	-233.65(7.90)	-416.56(11.79)	-777.85(15.65)	-1862.05(28.29)
	EM	-233.65(7.90)	-416.56(11.79)	-777.85(15.65)	-1862.05(28.29)
	FS	-233.65(7.90)	-416.56(11.79)	-777.85(15.65)	-1862.05(28.29)
	lme4	-233.65(7.90)	-416.56(11.79)	-777.85(15.65)	-1862.05(28.29)
20.00	MM	-242.21(8.20)	-424.68(10.11)	-795.63(15.56)	-1864.07(24.56)
	EM	-242.21(8.20)	-424.68(10.11)	-795.63(15.56)	-1864.07(24.56)
	FS	-242.21(8.20)	-424.68(10.11)	-795.63(15.56)	-1864.07(24.56)
	lme4	-242.21(8.20)	-424.68(10.11)	-795.63(15.56)	-1864.07(24.56)

To prove the assertion under Assumption 2, consider first the case $\boldsymbol{V}_1 = \boldsymbol{I}_n$. Setting $\alpha_i = \sigma_i^2/\sigma_1^2$ for $i = 2, \ldots, m$ reduces the loglikelihood to

$$L(\sigma_1^2, \boldsymbol{\alpha}) \;=\; -\frac{n}{2}\ln\sigma_1^2 - \frac{1}{2}\ln\det\Big(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\Big) - \frac{1}{2\sigma_1^2}\boldsymbol{y}^T\Big(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\Big)^{-1}\boldsymbol{y}. \quad (1)$$

The middle term on the right satisfies

$$-\frac{1}{2}\ln\det\Big(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\Big) \;\le\; 0$$

because $\det\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\right) \ge \det \boldsymbol{I}_n = 1$. Now let $\boldsymbol{U} = (\boldsymbol{U}_q, \boldsymbol{U}_{n-q})$ be an $n \times n$ orthogonal matrix whose left columns $\boldsymbol{U}_q$ span $\mathcal{H}$ and whose right columns $\boldsymbol{U}_{n-q}$ span $\mathcal{H}^\perp$. The identity

$$\boldsymbol{U}^T\Big(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\Big)\boldsymbol{U} \;=\; \begin{pmatrix} \boldsymbol{I}_q + \sum_{i=2}^m \alpha_i \boldsymbol{U}_q^T \boldsymbol{V}_i \boldsymbol{U}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-q} \end{pmatrix}$$

follows from the orthogonality relations $\boldsymbol{U}_{n-q}^T\boldsymbol{V}_i = \boldsymbol{U}_{n-q}^T\boldsymbol{U}_q = \boldsymbol{0}_{(n-q)\times n}$. This in turn implies

$$\begin{aligned}\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\right)^{-1} &= \boldsymbol{U}\begin{pmatrix}(\boldsymbol{I}_q + \sum_{i=2}^m \alpha_i \boldsymbol{U}_q^T\boldsymbol{V}_i\boldsymbol{U}_q)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-q}\end{pmatrix}\boldsymbol{U}^T \\ &\succeq \boldsymbol{U}\begin{pmatrix}\boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-q}\end{pmatrix}\boldsymbol{U}^T \\ &= \boldsymbol{U}_{n-q}\boldsymbol{U}_{n-q}^T.\end{aligned}$$

Therefore the quadratic term in equation (1) is bounded below by the positive constant

$$\boldsymbol{y}^T\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i \boldsymbol{V}_i\right)^{-1}\boldsymbol{y} \;\geq\; \boldsymbol{y}^T\boldsymbol{U}_{n-q}\boldsymbol{U}_{n-q}^T\boldsymbol{y} \;=\; \|\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y}\|^2 \;>\; 0.$$

Here the assumption $\boldsymbol{y} \notin \mathcal{H}$ guarantees the projection property $\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y} \neq \boldsymbol{0}$.

Next we show that the loglikelihood tends to $-\infty$ when σ_1^2 tends to 0 or ∞ or when $\|\boldsymbol{\alpha}\|_2$ tends to ∞. The second of the two inequalities

$$\begin{aligned}L(\sigma_0^2, \boldsymbol{\alpha}) &\leq -\frac{n}{2}\ln\sigma_1^2 - \frac{1}{2}\ln\det\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i\boldsymbol{V}_i\right) - \frac{1}{2\sigma_1^2}\|\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y}\|^2 \\ &\leq -\frac{n}{2}\ln\sigma_1^2 - \frac{1}{2\sigma_1^2}\|\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y}\|^2\end{aligned}$$

renders the claim about σ_1^2 obvious. To prove the claim about $\boldsymbol{\alpha}$, we make the worst case choice $\sigma_i^2 = \|\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y}\|^2$ in the first inequality. It follows that

$$L(\sigma_0^2, \boldsymbol{\alpha}) \;\leq\; -\frac{1}{2}\ln\det\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i\boldsymbol{V}_i\right) - \frac{n}{2}\ln\|\mathbb{P}_{\mathcal{H}^\perp}\boldsymbol{y}\|^2 - \frac{n}{2}.$$

If α_j tends to ∞, then the inequality

$$-\frac{1}{2}\ln\det\left(\boldsymbol{I}_n + \sum_{i=2}^m \alpha_i\boldsymbol{V}_i\right) \;\leq\; -\frac{1}{2}\ln\det\left(\boldsymbol{I}_n + \alpha_j\boldsymbol{V}_j\right) \;=\; -\frac{1}{2}\sum_{k=1}^n \ln(1 + \alpha_j\lambda_{jk})$$

holds, where the λ_{jk} are the eigenvalues of $\boldsymbol{V}_j$. At least one of these eigenvalues is positive because $\boldsymbol{V}_j$ is nontrivial. It follows that $L(\sigma_0^2, \boldsymbol{\alpha})$ tends to $-\infty$ in this case as well.

For the general case where $\boldsymbol{V}_1$ is non-singular but not necessarily $\boldsymbol{I}_n$, let $\boldsymbol{V}_1^{1/2}$ be the symmetric square root of $\boldsymbol{V}_1$ and write

$$\boldsymbol{V}_1 + \sum_{i=2}^m \sigma_i^2\boldsymbol{V}_i = \boldsymbol{V}_1^{1/2}\left(\boldsymbol{I} + \sum_{i=2}^m \sigma_i^2\boldsymbol{V}_1^{-1/2}\boldsymbol{V}_i\boldsymbol{V}_1^{-1/2}\right)\boldsymbol{V}_1^{1/2}.$$

The above arguments still apply since each $\boldsymbol{V}_1^{-1/2}\boldsymbol{V}_i\boldsymbol{V}_1^{-1/2}$ is nontrivial and $\boldsymbol{y}$ belongs to the $\text{span}\{\boldsymbol{V}_2, \ldots, \boldsymbol{V}_m\} = \boldsymbol{S}$ if and only if $\boldsymbol{V}_1^{-1/2}\boldsymbol{y}$ belongs to $\boldsymbol{V}_1^{-1/2}\boldsymbol{S}\boldsymbol{V}_1^{-1/2}$. □

Lemma S.2. *The iterates possess the ascent property* $L(M(\boldsymbol{\sigma}^{2(t)})) \geq L(\boldsymbol{\sigma}^{2(t)})$. *Furthermore, when* $L(M(\boldsymbol{\sigma}_*^2)) = L(\boldsymbol{\sigma}_*^2)$, $\boldsymbol{\sigma}_*^2$ *fulfills the fixed point condition* $M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$, *and each component satisfies either (i)* $\sigma_{*i}^2 = 0$ *or (ii)* $\sigma_{*i}^2 > 0$ *and* $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}_*^2) = 0$.

Proof. The ascent property is built into any MM algorithm. Suppose $L(M(\boldsymbol{\sigma}_*^2)) = L(\boldsymbol{\sigma}_*^2)$ at a point $\boldsymbol{\sigma}_*^2 \in \mathbb{R}_+^m$. Then equality must hold in the string of inequalities (3). It follows that

$$g(M(\boldsymbol{\sigma}_*^2) \mid \boldsymbol{\sigma}_*^2) \;=\; g(\boldsymbol{\sigma}_*^2 \mid \boldsymbol{\sigma}_*^2).$$

$g(\cdot \mid \boldsymbol{\sigma}_*^2)$ has a unique maximum since its Hessian is diagonal with strictly negative entries, hence $M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$. If $\sigma_{*i}^2 > 0$, the stationarity condition

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}_*^2) \;=\; \frac{\partial}{\partial \sigma_i^2} g(\boldsymbol{\sigma}_*^2 \mid \boldsymbol{\sigma}_*^2) \;=\; 0$$

applies. The equivalence of the two displayed partial derivatives is a consequence of the fact that the difference $f(\boldsymbol{\sigma}^2) - g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}_*^2)$ achieves its minimum of 0 at $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}_*^2$. □

Lemma S.3. *The distance between successive iterates* $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2$ *converges to 0.*

Proof. Suppose on the contrary that $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2$ does not converge to 0. Then one can extract a subsequence $\{t_k\}_{k \geq 1}$ such that

$$\|\boldsymbol{\sigma}^{2(t_k+1)} - \boldsymbol{\sigma}^{2(t_k)}\|_2 \geq \epsilon > 0 \tag{2}$$

for all k. Let C_0 be the compact super-level set $\{\boldsymbol{\sigma}^2 : L(\boldsymbol{\sigma}^2) \geq L(\boldsymbol{\sigma}^{2(0)})\}$. Since the sequence $\{\boldsymbol{\sigma}^{2(t_k)}\}_{k \geq 1}$ is confined to C_0, one can pass to a subsequence if necessary and assume that $\boldsymbol{\sigma}^{2(t_k)}$ converges to a limit $\boldsymbol{\sigma}_*^2$ and that $\boldsymbol{\sigma}^{2(t_k+1)}$ converges to a limit $\boldsymbol{\sigma}_{**}^2$. Taking limits in the relation $\boldsymbol{\sigma}^{2(t_k+1)} = M(\boldsymbol{\sigma}^{2(t_k)})$ and invoking the continuity $M(\boldsymbol{\sigma}^2)$ imply that $\boldsymbol{\sigma}_{**}^2 = M(\boldsymbol{\sigma}_*^2)$. Because the sequence $L(\boldsymbol{\sigma}^{2(t_k)})$ is monotonically increasing in k and bounded above on C_0, it converges to a limit L_*. Hence, the continuity of $L(\boldsymbol{\sigma}^2)$ implies

$$L(\boldsymbol{\sigma}_*^2) \;=\; \lim_k L(\boldsymbol{\sigma}^{2(t_k)}) \;=\; L_* \;=\; \lim_k L(\boldsymbol{\sigma}^{2(t_k+1)}) \;=\; L(\boldsymbol{\sigma}_{**}^2) \;=\; L(M(\boldsymbol{\sigma}_*^2)).$$

Lemma S.2 therefore gives $\boldsymbol{\sigma}_{**}^2 = M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$, contradicting the bound $\|\boldsymbol{\sigma}_*^2 - \boldsymbol{\sigma}_{**}^2\|_2 \geq \epsilon$ entailed by inequality (2). □

With Lemmas S.1-S.3, we are ready to prove Theorem 1.

Proof. The sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t \geq 0}$ is contained in the super-level compact set C_0 defined in Lemma S.3 and therefore admits a convergent subsequence $\boldsymbol{\sigma}^{2(t_k)}$ with limit $\boldsymbol{\sigma}^{2(\infty)}$. As argued in Lemma S.3, $L(\boldsymbol{\sigma}^{2(\infty)}) = L(M(\boldsymbol{\sigma}^{2(\infty)}))$. Lemma S.2 now implies that $\boldsymbol{\sigma}^{2(\infty)}$ is a fixed point of the algorithm map $M(\boldsymbol{\sigma}^2)$.

According to Ostrowski's theorem (Lange, 2010, Proposition 8.2.1), the set of limit points of a bounded sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t \geq 0}$ is connected and compact provided $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2 \to 0$. If the set of fixed points is discrete, then the connected subset of limit points reduces to a single point. Hence, the bounded sequence $\boldsymbol{\sigma}^{2(t)}$ converges to this point. When the limit exists, one can check that $\boldsymbol{\sigma}^{2(\infty)}$ satisfies the KKT conditions by proving that each zero component of $\boldsymbol{\sigma}^{2(\infty)}$ has a non-positive partial derivative. Suppose on the contrary $\sigma_i^{2(\infty)} = 0$ and $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^{2(\infty)}) > 0$. By continuity $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^{2(t)}) > 0$ for all large t. Therefore, $\sigma_i^{2(t+1)} > \sigma_i^{2(t)}$ for all large t by the observation made after equation (9). This behavior is inconsistent with the assumption that $\sigma_i^{2(t)} \to 0$. □

S.4 Proof of Theorem 2

MM algorithm: The minorizing function for the MM algorithm is

$$\begin{aligned} & g_{\text{MM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) \\ = & -\frac{1}{2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{\Omega}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\right)\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)}) + c^{(t)} \\ = & \sum_{i=1}^m -\frac{\sigma_i^2}{2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i) - \frac{\sigma_i^{4(t)}}{2\sigma_i^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)}) + c^{(t)}, \end{aligned}$$

where

$$c^{(t)} = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det\boldsymbol{\Omega}^{(t)} + \frac{n}{2}.$$

Taking derivatives, we have

$$\begin{aligned} \frac{\partial}{\partial\sigma_i^2}g_{\text{MM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) &= -\frac{1}{2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i) + \frac{\sigma_i^{4(t)}}{2\sigma_i^4}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)}), \\ \frac{\partial^2}{\partial\sigma_i^2\partial\sigma_j^2}g_{\text{MM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) &= \begin{cases} -\frac{\sigma_i^{4(t)}}{\sigma_i^6}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)}) & i=j \\ 0 & i\neq j. \end{cases} \end{aligned}$$

EM algorithm: Assume $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^m \boldsymbol{Z}_i$, where $\boldsymbol{Z}_i \sim N(\boldsymbol{0}, \sigma_i^2\boldsymbol{V}_i)$ are independent. Then the complete data is $\boldsymbol{Z} = (\boldsymbol{Z}_i, \cdots, \boldsymbol{Z}_m)$. From the information inequality, we have

$$L(\boldsymbol{y}|\boldsymbol{\sigma}^2) \;\geq\; Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) - Q(\boldsymbol{\sigma}^{2(t)}|\boldsymbol{\sigma}^{2(t)}) + L(\boldsymbol{y}|\boldsymbol{\sigma}^{2(t)}),$$

where

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\sigma}^2) &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}), \\ Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) &= -\frac{1}{2}\sum_{i=1}^m\left[\text{rank}(\boldsymbol{V}_i)\ln\sigma_i^2 + \frac{\sigma_i^{2(t)}}{\sigma_i^2}\text{rank}(\boldsymbol{V}_i) - \frac{\sigma_i^{4(t)}}{\sigma_i^2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)\right] \\ &\quad -\frac{1}{2}\sum_{i=1}^m\left[\frac{\sigma_i^{4(t)}}{\sigma_i^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})\right]. \end{aligned}$$

We derive this expression in Section S.6. The minorizing function

$$\begin{aligned} & g_{\text{EM}}(\boldsymbol{\sigma}|\boldsymbol{\sigma}^{2(t)}) \\ = & Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) - Q(\boldsymbol{\sigma}^{2(t)}|\boldsymbol{\sigma}^{2(t)}) + L(\boldsymbol{y}|\boldsymbol{\sigma}^{2(t)}) \\ = & -\frac{1}{2}\sum_{i=1}^m\left[\text{rank}(\boldsymbol{V}_i)\ln\sigma_i^2 + \frac{\sigma_i^{2(t)}}{\sigma_i^2}\text{rank}(\boldsymbol{V}_i) - \frac{\sigma_i^{4(t)}}{\sigma_i^2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)\right] \\ & -\frac{1}{2}\sum_{i=1}^m\left[\frac{\sigma_i^{4(t)}}{\sigma_i^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})\right] \\ & -\frac{1}{2}\sum_{i=1}^m\left[-\text{rank}(\boldsymbol{V}_i)\ln\sigma_i^{2(t)} - \text{rank}(\boldsymbol{V}_i)\right] - \frac{}{2}\left[n\ln(2\pi) + \ln\det\boldsymbol{\Omega}^{(t)} + n\right] \end{aligned}$$

of the EM algorithms depends on $\boldsymbol{\sigma}^2$ only through $Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)})$. Taking derivatives, we have

$$\begin{aligned}
&\frac{\partial}{\partial \sigma_i^2} g_{\text{EM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) \\
= \;& -\frac{\text{rank}(\boldsymbol{V}_i)}{2\sigma_i^2} + \frac{\text{rank}(\boldsymbol{V}_i)\sigma_i^{2(t)} - \sigma_i^{4(t)}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i) + \sigma_i^{4(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})}{2\sigma_i^4}, \\
&\frac{\partial^2}{\partial \sigma_i^2 \partial \sigma_j^2} g_{\text{EM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}) \\
= \;& \begin{cases} \frac{\text{rank}(\boldsymbol{V}_i)}{2\sigma_i^4} - \frac{\text{rank}(\boldsymbol{V}_i)\sigma_i^{2(t)} - \sigma_i^{4(t)}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i) + \sigma_i^{4(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})}{\sigma_i^6} & i = j \\ 0 & i \neq j. \end{cases}
\end{aligned}$$

EM vs MM: Let $\boldsymbol{\sigma}^{2(\infty)}$ be a common limit point of EM and MM. By Lemma S.2, each component of $\boldsymbol{\sigma}^{2(\infty)}$ is either 0 or has vanishing gradient. Therefore

$$\begin{aligned}
\frac{\partial^2}{(\partial \sigma_i^2)^2} g_{\text{EM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(\infty)}) \mid_{\boldsymbol{\sigma}^2=\boldsymbol{\sigma}^{2(\infty)}} &= -\frac{\text{rank}(\boldsymbol{V}_i)}{2\sigma_i^{4(\infty)}}, \\
\frac{\partial^2}{(\partial \sigma_i^2)^2} g_{\text{MM}}(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(\infty)}) \mid_{\boldsymbol{\sigma}^2=\boldsymbol{\sigma}^{2(\infty)}} &= -\frac{\text{tr}(\boldsymbol{\Omega}^{-(\infty)}\boldsymbol{V}_i)}{\sigma_i^{2(\infty)}}
\end{aligned}$$

and, when all the $\boldsymbol{V}_i$ all non-singular,

$$\begin{aligned}
& \frac{1}{m}\sum_{i=1}^m \frac{[d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)}|\boldsymbol{\sigma}^{2(\infty)})]_{ii}}{[d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)}|\boldsymbol{\sigma}^{2(\infty)})]_{ii}} \\
= \;& \frac{2}{m}\sum_{i=1}^m \frac{\sigma_i^{2(\infty)}\text{tr}(\boldsymbol{\Omega}^{-(\infty)}\boldsymbol{V}_i)}{\text{rank}(\boldsymbol{V}_i)} \\
= \;& \frac{2}{m} \;\leq\; 1.
\end{aligned}$$

S.5 Derivation of Algorithm 4

When there are $m = 2$ variance components $\boldsymbol{\Omega} = \boldsymbol{\Gamma}_1 \otimes \boldsymbol{V}_1 + \boldsymbol{\Gamma}_2 \otimes \boldsymbol{V}_2$, repeated inversion of the $nd \times nd$ covariance matrix $\boldsymbol{\Omega}$ can be reduced to one $d \times d$ (generalized) eigen-decomposition per iteration. The generalized eigen-decomposition of the matrix pair $(\boldsymbol{V}_1, \boldsymbol{V}_2)$ yields generalized eigenvalues $\boldsymbol{d} = (d_1, \ldots, d_n)^T$ and generalized eigenvectors $\boldsymbol{U}$ such that $\boldsymbol{U}^T\boldsymbol{V}_1\boldsymbol{U} = \boldsymbol{D} = \text{diag}(\boldsymbol{d})$ and $\boldsymbol{U}^T\boldsymbol{V}_2\boldsymbol{U} = \boldsymbol{I}$. Let the generalized eigen-decomposition of $(\boldsymbol{\Gamma}_1^{(t)}, \boldsymbol{\Gamma}_2^{(t)})$ be $(\boldsymbol{\Lambda}^{(t)}, \boldsymbol{\Phi}^{(t)})$ such that $\boldsymbol{\Phi}^{(t)T}\boldsymbol{\Gamma}_1^{(t)}\boldsymbol{\Phi}^{(t)} = \boldsymbol{\Lambda}^{(t)} = \text{diag}(\boldsymbol{\lambda}^{(t)})$ and $\boldsymbol{\Phi}^{(t)T}\boldsymbol{\Gamma}_2\boldsymbol{\Phi}^{(t)} = \boldsymbol{I}_d$. It follows that

$$\begin{aligned}
\boldsymbol{\Omega}^{(t)} &= (\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1})^T(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n)(\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1}) \\
\boldsymbol{\Omega}^{-(t)} &= (\boldsymbol{\Phi}^{(t)} \otimes \boldsymbol{U})(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n)^{-1}(\boldsymbol{\Phi}^{(t)} \otimes \boldsymbol{U})^T \\
\det \boldsymbol{\Omega}^{(t)} &= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \det(\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1})^T(\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1}) \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \det(\boldsymbol{\Gamma}_2^{(t)} \otimes \boldsymbol{V}_2) \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \det(\boldsymbol{\Gamma}_2^{(t)})^n \det(\boldsymbol{V}_2)^d.
\end{aligned}$$

To update the fixed effects $\boldsymbol{B}$ given $\boldsymbol{\Gamma}_1^{(t)}$ and $\boldsymbol{\Gamma}_2^{(t)}$, the general least squares criterion is

$$\begin{aligned}
& \frac{1}{2}[\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})]^T\boldsymbol{\Omega}^{-(t)}[\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})] \\
= & \frac{1}{2}[\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})]^T(\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1}(\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})^T[\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})] \\
= & \frac{1}{2}\mathrm{vec}[\boldsymbol{U}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Phi}^{(t)}]^T(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1}\mathrm{vec}[\boldsymbol{U}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Phi}^{(t)}] \\
= & \frac{1}{2}[\mathrm{vec}(\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{\Phi}^{(t)})-(\boldsymbol{\Phi}^{(t)T}\otimes\boldsymbol{U}^T\boldsymbol{X})\mathrm{vec}\boldsymbol{B}]^T(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1} \\
& \cdot[\mathrm{vec}(\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{\Phi}^{(t)})-(\boldsymbol{\Phi}^{(t)T}\otimes\boldsymbol{U}^T\boldsymbol{X})\mathrm{vec}\boldsymbol{B}].
\end{aligned}$$

Minimization of this criterion reduces to a weighted least squares problem for the transformed responses $\tilde{\boldsymbol{Y}}=\boldsymbol{U}^T\boldsymbol{Y}$, transformed predictor matrix $\tilde{\boldsymbol{X}}=\boldsymbol{U}^T\boldsymbol{X}$, and observation weights $(\lambda_k^{(t)}d_i+1)^{-1}$. To update $\boldsymbol{\Gamma}_1^{(t)}$ and $\boldsymbol{\Gamma}_2^{(t)}$, we need to evaluate the matrices $\boldsymbol{M}_i$ and $\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}$ that appear in the stationarity condition (13).

Evaluation of $\boldsymbol{M}_i$: Note the (j,k)-th entry of $\boldsymbol{M}_i$ is $\mathrm{tr}(\boldsymbol{\Omega}_{jk}^{-(t)}\boldsymbol{V}_i)$, where $\boldsymbol{\Omega}_{jk}^{-(t)}$ is the (j,k)-th block of

$$\boldsymbol{\Omega}^{-(t)} = (\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1}(\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})^T,$$

which can be expressed as

$$\boldsymbol{\Omega}_{jk}^{-(t)} = \sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\boldsymbol{U}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\boldsymbol{U}^T.$$

Therefore $\boldsymbol{M}_1$ has entries

$$\begin{aligned}
(\boldsymbol{M}_1)_{jk} &= \mathrm{tr}(\boldsymbol{V}_1\boldsymbol{\Omega}_{ij}^{-(t)}) \\
&= \mathrm{tr}\Big[\boldsymbol{U}^{-T}\boldsymbol{D}\boldsymbol{U}^{-1}\sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\boldsymbol{U}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\boldsymbol{U}^T\Big] \\
&= \mathrm{tr}\Big[\sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\boldsymbol{D}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\Big] \\
&= \sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\mathrm{tr}\Big[\boldsymbol{D}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\Big],
\end{aligned}$$

and $\boldsymbol{M}_2$ has entries

$$\begin{aligned}
(\boldsymbol{M}_2)_{jk} &= \mathrm{tr}(\boldsymbol{V}_2\boldsymbol{\Omega}_{ij}^{-(t)}) \\
&= \mathrm{tr}\Big[\boldsymbol{U}^{-T}\boldsymbol{U}^{-1}\sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\boldsymbol{U}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\boldsymbol{U}^T\Big] \\
&= \mathrm{tr}\Big[\sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\Big] \\
&= \sum_{l=1}^{d}\phi_{jl}^{(t)}\phi_{lk}^{(t)}\mathrm{tr}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}.
\end{aligned}$$

Collectively we have

$$\begin{aligned}\boldsymbol{M}_1 &= \boldsymbol{\Phi}^{(t)}\mathrm{diag}\Big\{\mathrm{tr}\Big[\boldsymbol{D}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\Big]\Big\}\boldsymbol{\Phi}^{(t)T}\\ \boldsymbol{M}_2 &= \boldsymbol{\Phi}^{(t)}\mathrm{diag}\Big[\mathrm{tr}(\lambda_l^{(t)}\boldsymbol{D}+\boldsymbol{I}_n)^{-1}\Big]\boldsymbol{\Phi}^{(t)T}.\end{aligned}$$

Evaluation of $\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}$: Write

$$\begin{aligned}\boldsymbol{\Gamma}_1^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_1\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_1^{(t)} &= \boldsymbol{N}_1^T\boldsymbol{N}_1\\ \boldsymbol{\Gamma}_2^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_2\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_2^{(t)} &= \boldsymbol{N}_2^T\boldsymbol{N}_2,\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{N}_1 &= \boldsymbol{D}^{1/2}\boldsymbol{U}^{-1}\boldsymbol{R}^{(t)}\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}\\ \boldsymbol{N}_2 &= \boldsymbol{U}^{-1}\boldsymbol{R}^{(t)}\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Phi}^{-(t)}.\end{aligned}$$

To further simplify, note

$$\begin{aligned}&\mathrm{vec}\,\boldsymbol{N}_1\\ =\;&(\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}\otimes\boldsymbol{D}^{1/2}\boldsymbol{U}^{-1})\mathrm{vec}\boldsymbol{R}^{(t)}\\ =\;&(\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}\otimes\boldsymbol{D}^{1/2}\boldsymbol{U}^{-1})\boldsymbol{\Omega}^{-(t)}\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})\\ =\;&(\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}\otimes\boldsymbol{D}^{1/2}\boldsymbol{U}^{-1})(\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1}(\boldsymbol{\Phi}^{(t)}\otimes\boldsymbol{U})^T\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})\\ =\;&(\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}^{1/2})(\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}+\boldsymbol{I}_d\otimes\boldsymbol{I}_n)^{-1}\mathrm{vec}(\boldsymbol{U}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)})\\ =\;&(\boldsymbol{\Phi}^{-(t)T}\boldsymbol{\Lambda}^{(t)}\otimes\boldsymbol{D}^{1/2})\mathrm{vec}[\boldsymbol{U}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)}\oslash(\boldsymbol{d}\boldsymbol{\lambda}^{(t)T}+\mathbf{1}_n\mathbf{1}_d^T)]\\ =\;&\mathrm{vec}\,\{\boldsymbol{D}^{1/2}[(\boldsymbol{U}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)})\oslash(\boldsymbol{d}\boldsymbol{\lambda}^{(t)T}+\mathbf{1}_n\mathbf{1}_d^T)]\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}\},\end{aligned}$$

where $\oslash$ denotes a Hadamard quotient. Thus,

$$\boldsymbol{N}_1 = \boldsymbol{D}^{1/2}\{[(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)}]\oslash(\boldsymbol{d}\boldsymbol{\lambda}^{(t)T}+\mathbf{1}_n\mathbf{1}_d^T)\}\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)},$$

and similarly

$$\boldsymbol{N}_2 = \{[(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)}]\oslash(\boldsymbol{d}\boldsymbol{\lambda}^{(t)T}+\mathbf{1}_n\mathbf{1}_d^T)\}\boldsymbol{\Phi}^{-(t)}.$$

S.6 EM Algorithm for the Multivariate Response Model

In this section we review the derivation of the EM algorithm for the multivariate response model (Glanz and Carvalho, 2013; Reinsel, 1984). If the response matrix $\boldsymbol{Y}$ can be written as the sum $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{B}+\boldsymbol{Z}_1+\cdots+\boldsymbol{Z}_m$ of independent random matrices with $\mathrm{vec}\,\boldsymbol{Z}_i\sim N(\mathbf{0},\boldsymbol{\Omega}_i)$, then $\mathrm{vec}\,\boldsymbol{Y}\sim N(\mathrm{vec}(\boldsymbol{X}\boldsymbol{\beta}),\boldsymbol{\Omega})$, where $\boldsymbol{\Omega}=\sum_{i=1}^m\boldsymbol{\Omega}_i$. Under the matrix normal assumption, $\boldsymbol{\Omega}_i=\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i$. As in the text, the $p\times d$ coefficient matrix $\boldsymbol{B}$ collects the fixed effects, the $\boldsymbol{\Gamma}_i$ are unknown $d\times d$ covariance matrices, and the $\boldsymbol{V}_i$ are known $n\times n$ covariance matrices. The complete data log-likelihood for the unobserved $\boldsymbol{Z}_i$ is

$$-\frac{1}{2}\sum_{i=1}^m\ln\det{}^+\boldsymbol{\Omega}_i-\frac{1}{2}\sum_{i=1}^m,\mathrm{vec}(\boldsymbol{Z}_i)^T\boldsymbol{\Omega}_i^+\mathrm{vec}(\boldsymbol{Z}_i),$$

where $\det^+ \boldsymbol{\Omega}_i$ denotes the pseudo-determinant of $\boldsymbol{\Omega}_i$ and $\boldsymbol{\Omega}_i^+$ the pseudo-inverse of $\boldsymbol{\Omega}_i$. To compute the surrogate function for the EM algorithm, one needs the conditional expectations

$$\mathrm{E}(\mathrm{vec}\,\boldsymbol{Z}_i \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(t)}) \;=\; \boldsymbol{\Omega}_i^{(t)} \boldsymbol{\Omega}^{-(t)} \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t)}) \;=\; \boldsymbol{E}_i^{(t)}$$

and the conditional covariances

$$\mathrm{Cov}(\mathrm{vec}\,\boldsymbol{Z}_i \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(t)}) \;=\; \boldsymbol{\Omega}_i^{(t)} - \boldsymbol{\Omega}_i^{(t)} \boldsymbol{\Omega}^{-(t)} \boldsymbol{\Omega}_i^{(t)} \;=\; \boldsymbol{F}_i^{(t)},$$

where $\boldsymbol{\theta}$ is the parameter vector. These are employed to compute the conditional second moments

$$\mathrm{E}(\mathrm{vec}\,\boldsymbol{Z}_i\,\mathrm{vec}\,\boldsymbol{Z}_i^T \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(t)}) \;=\; \boldsymbol{F}_i^{(t)} + \boldsymbol{E}_i^{(t)} (\boldsymbol{E}_i^{(t)})^T \;=\; \boldsymbol{G}_i^{(t)}.$$

Here the random vector $\boldsymbol{Z}_i$ should be replaced by $\boldsymbol{Z}_m - \boldsymbol{X}\boldsymbol{B}^{(t)}$ when $i = m$.

One can readily check that $\boldsymbol{\Omega}_i^+ = \boldsymbol{\Gamma}_i^+ \otimes \boldsymbol{V}_i^+ = \boldsymbol{\Gamma}_i^{-1} \otimes \boldsymbol{V}_i^+$ for $\boldsymbol{\Gamma}_i$ invertible. Since the pseudo-determinant of a positive semidefinite matrix equals the product of its positive eigenvalues, the formulas

$$\begin{aligned} \det{}^+ \boldsymbol{\Omega}_i &= (\det\, \boldsymbol{\Gamma}_i)^{r_i} (\det{}^+ \boldsymbol{V}_i)^{s_i} \\ \ln \det{}^+ \boldsymbol{\Omega}_i &= r_i \ln \det \boldsymbol{\Gamma}_i + s_i \ln \det{}^+ \boldsymbol{V}_i \end{aligned}$$

apply, where $r_i = \mathrm{rank}(\boldsymbol{V}_i^+)$ and $s_i = \mathrm{rank}(\boldsymbol{\Gamma}_i^+)$. In the M step of the EM algorithm, one maximizes the surrogate

$$-\frac{1}{2} \sum_{i=1}^m r_i \ln \det\, \boldsymbol{\Gamma}_i - \frac{1}{2} \sum_{i=1}^m \mathrm{tr}[(\boldsymbol{\Gamma}_i^{-1} \otimes \boldsymbol{V}_i^+) \boldsymbol{G}_i^{(t)}]. \tag{3}$$

For $\boldsymbol{\Gamma}_i$ unstructured, we substitute $\boldsymbol{\Lambda}_i = \boldsymbol{\Gamma}_i^{-1}$ and maximize with respect to $\boldsymbol{\Lambda}_i$. Fortunately, the next lemma can be invoked.

Lemma S.4. *If the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are $d \times d$, $n \times n$, and $dn \times dn$ respectively, then*

$$\mathrm{tr}[(\boldsymbol{A} \otimes \boldsymbol{B}) \boldsymbol{C}^T] \;=\; \mathrm{tr}\{(\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T [(\boldsymbol{1}_d \boldsymbol{1}_d^T \otimes \boldsymbol{B}) \odot \boldsymbol{C}] (\boldsymbol{I}_d \otimes \boldsymbol{1}_n) \boldsymbol{A}^T\}.$$

Proof. This trace identity is essentially proved in the text. □

Lemma S.4 yields

$$\mathrm{tr}[(\boldsymbol{\Lambda}_i \otimes \boldsymbol{V}_i^+) \boldsymbol{G}_i^{(t)}] \;=\; \mathrm{tr}\{(\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T [(\boldsymbol{1}_d \boldsymbol{1}_d^T \otimes \boldsymbol{V}_i^+) \odot \boldsymbol{G}_i^{(t)}] (\boldsymbol{I}_d \otimes \boldsymbol{1}_n) \boldsymbol{\Lambda}_i\}.$$

The stationarity condition

$$\boldsymbol{0} \;=\; \frac{1}{2} r_i \boldsymbol{\Lambda}_i^{-1} - \frac{1}{2} (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T [(\boldsymbol{1}_d \boldsymbol{1}_d^T \otimes \boldsymbol{V}_i^+) \odot \boldsymbol{G}_i^{(t)}] (\boldsymbol{I}_d \otimes \boldsymbol{1}_n).$$

now entails the update

$$\boldsymbol{\Gamma}_i^{(t+1)} \;=\; \frac{1}{r_i} (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^T [(\boldsymbol{1}_d \boldsymbol{1}_d^T \otimes \boldsymbol{V}_i^+) \odot \boldsymbol{G}_i^{(t)}] (\boldsymbol{I}_d \otimes \boldsymbol{1}_n).$$

In the case $m = 1$, the single update reduces to

$$\boldsymbol{\Gamma}^{(t+1)} = \frac{1}{r}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t)})^T \boldsymbol{V}^{+}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t)}),$$

which matches the earlier result of Glanz and Carvalho (2013). When $\boldsymbol{\Gamma}_i$ is the scalar σ_i^2,

$$\begin{aligned}\boldsymbol{E}_i^{(t)} &= \sigma_i^{2(t)} \boldsymbol{V}_i \boldsymbol{\Omega}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}) \\ \boldsymbol{F}_i^{(t)} &= \sigma_i^{2(t)} \boldsymbol{V}_i - \sigma_i^{2(t)} \boldsymbol{V}_i \boldsymbol{\Omega}^{-(t)} \sigma_i^{2(t)} \boldsymbol{V}_i.\end{aligned}$$

One recovers the representation (3) by substituting these quantities in equation (3) and invoking the identities $\boldsymbol{V}_i \boldsymbol{V}_i^{+} \boldsymbol{V}_i = \boldsymbol{V}_i$ and $\mathrm{tr}(\boldsymbol{V}_i \boldsymbol{V}_i^{+}) = \mathrm{rank}(\boldsymbol{V}_i)$ and the cyclic permutation property of the trace.

S.7 Proof of Lemma 3

Proof. Direct substitution shows that $\boldsymbol{Y}$ solves the equivalent equation $\boldsymbol{X}\boldsymbol{B}\boldsymbol{X} = \boldsymbol{A}$. To show uniqueness, suppose $\boldsymbol{Y}^{-1}\boldsymbol{A}\boldsymbol{Y}^{-1} = \boldsymbol{B}$ and $\boldsymbol{Z}^{-1}\boldsymbol{A}\boldsymbol{Z}^{-1} = \boldsymbol{B}$. The equations

$$\begin{aligned}(\boldsymbol{B}^{1/2}\boldsymbol{Y}\boldsymbol{B}^{1/2})^2 &= \boldsymbol{B}^{1/2}\boldsymbol{Y}\boldsymbol{B}\boldsymbol{Y}\boldsymbol{B}^{1/2} = \boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2} \\ (\boldsymbol{B}^{1/2}\boldsymbol{Z}\boldsymbol{B}^{1/2})^2 &= \boldsymbol{B}^{1/2}\boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}\boldsymbol{B}^{1/2} = \boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2}\end{aligned}$$

imply $\boldsymbol{B}^{1/2}\boldsymbol{Y}\boldsymbol{B}^{1/2} = \boldsymbol{B}^{1/2}\boldsymbol{Z}\boldsymbol{B}^{1/2}$ by virtue of the uniqueness of symmetric square root. Since $\boldsymbol{B}^{-1/2}$ is positive definite, $\boldsymbol{Y} = \boldsymbol{Z}$. □

S.8 Proof of Proposition 2

Proof. If $\boldsymbol{V}_i$ has strictly positive diagonal entries, then so does $\mathbf{1}_d\mathbf{1}_d^T \otimes \boldsymbol{V}_i$, and the Hadamard product $(\mathbf{1}_d\mathbf{1}_d^T \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}$ is positive definite by Schur's lemma. Since the matrix $\boldsymbol{I}_d \otimes \mathbf{1}_n$ has full column rank d, the matrix $\boldsymbol{M}_i$ is also positive definite. Finally, if no column of $\boldsymbol{R}^{(t)}$ lies in the null space of $\boldsymbol{V}_i$, and $\boldsymbol{\Gamma}^{(t)}$ is positive define, then $\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}$ is positive definite. The second claim follows by induction and Lemma 3. □

S.9 Proof of Lemma 4

Proof. Under the hypotheses, the representations $\boldsymbol{A}^{+} = (\boldsymbol{A}^T\boldsymbol{A})^{+}\boldsymbol{A}^T = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ and $\boldsymbol{B}^{+} = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}$ are well known. The choice $\boldsymbol{B}^{+}\boldsymbol{A}^{+} = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ satisfies the four equations characterizing the pseudo-inverse of $\boldsymbol{A}\boldsymbol{B}$. □

S.10 Proof of Lemma 5

Proof. Suppose $\boldsymbol{A}$ has spectral decomposition $\sum_i \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^T$. The matrix $\boldsymbol{P} = \sum_{\lambda_i > 0} \boldsymbol{u}_i\boldsymbol{u}_i^T$ projects onto the range of $\boldsymbol{A}$ and therefore also projects onto the range of $\boldsymbol{B}$. It follows that $\boldsymbol{P}\boldsymbol{B} = \boldsymbol{B}$ and by symmetry that $\boldsymbol{B}\boldsymbol{P} = \boldsymbol{B}$. This allows us to write

$$\begin{aligned}&(\boldsymbol{B} + \epsilon\boldsymbol{I})(\boldsymbol{A} + \epsilon\boldsymbol{I})^{-1}(\boldsymbol{B} + \epsilon\boldsymbol{I}) \\ = \;& \boldsymbol{B}\boldsymbol{P}(\boldsymbol{A} + \epsilon\boldsymbol{I})^{-1}\boldsymbol{P}\boldsymbol{B} + \epsilon\boldsymbol{B}\boldsymbol{P}(\boldsymbol{A} + \epsilon\boldsymbol{I})^{-1} + \epsilon(\boldsymbol{A} + \epsilon\boldsymbol{I})^{-1}\boldsymbol{P}\boldsymbol{B} + \epsilon^2(\boldsymbol{A} + \epsilon\boldsymbol{I})^{-1}.\end{aligned}$$

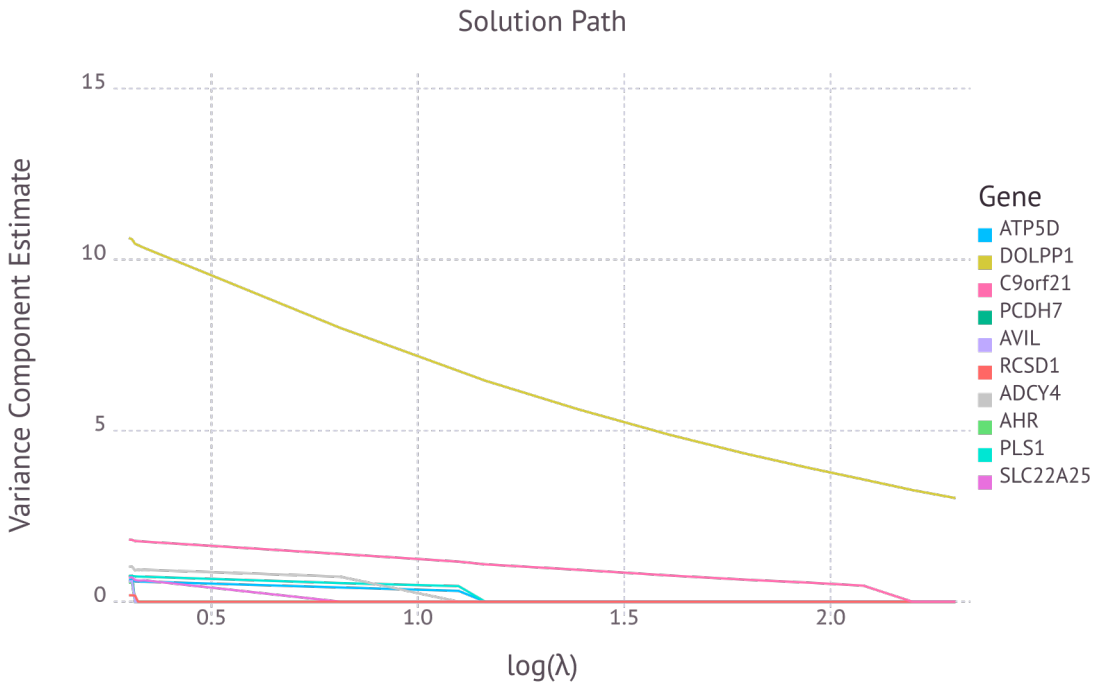

Figure 1: Solution path of the lasso penalized variance component model (17) indicates the top genes in an association study of 200 genes and the complex trait `height`.

The last three of these terms vanish as $\epsilon \downarrow 0$; the first term tends to the claimed limit. These assertions follow from the expressions

$$\boldsymbol{P}(\boldsymbol{A}+\epsilon\boldsymbol{I})^{-1}\boldsymbol{P} \;=\; \boldsymbol{P}(\boldsymbol{A}+\epsilon\boldsymbol{I})^{-1} \;=\; (\boldsymbol{A}+\epsilon\boldsymbol{I})^{-1}\boldsymbol{P} \;=\; \sum_{\lambda_i>0}\frac{1}{\lambda_i+\epsilon}\boldsymbol{u}_i\boldsymbol{u}_i^T$$

and $\epsilon^2(\boldsymbol{A}+\epsilon\boldsymbol{I})^{-1} = \sum_i \frac{\epsilon^2}{\lambda_i+\epsilon}\boldsymbol{u}_i\boldsymbol{u}_i^T$. □

S.11 Proof of Lemma 6

Proof. In fact, both matrices have range equal to the range of $\boldsymbol{Z}$. The matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}\boldsymbol{R}^{1/2}$ clearly have the same range. Furthermore, the matrices $\boldsymbol{Z}\boldsymbol{R}^{1/2}$ and $\boldsymbol{Z}\boldsymbol{R}^{1/2}\boldsymbol{R}^{1/2}\boldsymbol{Z}^T$ also have the same range. □

S.12 Lasso solution path

Figure 1 displays the lasso solution path for the QTL example in Section 7.

References

Glanz, H. and Carvalho, L. (2013). An expectation-maximization algorithm for the matrix normal distribution. *arXiv*, 1309.6609.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Statistics and Computing. Springer, New York, second edition.

Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.*, 79(386):406–414.