Sparse high-dimensional linear mixed modeling with a partitioned empirical Bayes ECM algorithm

SUPPLEMENTARY MATERIAL

Anja Zgodic[1], Ray Bai[2], Jiajia Zhang[1], Peter Olejua[1], Alexander C. McLain[1*]

[1*]Department of Epidemiology and Biostatistics, University of South Carolina, 915 Greene Street, Columbia, South Carolina, 29208, United States.

[2]Department of Statistics, University of South Carolina, 1523 Greene Street, Columbia, South Carolina, 29208, United States.

*Corresponding author(s). E-mail(s): mclaina@mailbox.sc.edu;

A Expanded Bayesian linear mixed model setup

We present some additional details relevant to the Bayesian linear mixed model framework.

Given σ^2, $\boldsymbol{\gamma}$, and $\boldsymbol{b}$, standard results can be leveraged to show that, when breaking $\boldsymbol{X}$ down into $(\boldsymbol{X}_\gamma, \boldsymbol{X}_{\bar{\gamma}})'$, for predictors where $\gamma_k = 1$ and $\gamma_k = 0$, respectively, and $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_\gamma\ \boldsymbol{\beta}_{\bar{\gamma}})'$, the conditional distribution of $\boldsymbol{\gamma\beta}$ is

$$\boldsymbol{\beta}_\gamma|(\boldsymbol{Y}, \sigma^2, \boldsymbol{\gamma}) \sim N\left\{\hat{\boldsymbol{\beta}}_\gamma, \sigma^2(\boldsymbol{X}'_\gamma \boldsymbol{X}_\gamma)^{-1}\right\}$$

for predictors $\boldsymbol{X}_\gamma$, where $\hat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}'_\gamma \boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}'_\gamma(\boldsymbol{Y} - \boldsymbol{A}'\boldsymbol{\omega} - \boldsymbol{\mathcal{V}}'\boldsymbol{b})$, while for predictors in $\boldsymbol{X}_{\bar{\gamma}}$ the conditional distribution of $\boldsymbol{\gamma\beta}$ is $\delta_0(\cdot)$, a point mass at zero. The conditional

distribution for σ^2 is inverse-gamma (IG) with parameters (a, d),

$$\sigma^2|(\boldsymbol{Y}, a, d, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\omega}) \sim IG\left\{\underbrace{\frac{M-2}{2}}_{a}, \underbrace{\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}_\gamma - \boldsymbol{A}'\boldsymbol{\omega} - \boldsymbol{\mathcal{V}}'\boldsymbol{b}\|_2^2}_{d}\right\},$$

where $\|\boldsymbol{x}\|_2$ denotes the ℓ_2-norm of $\boldsymbol{x}$. To obtain the conditional distribution of $\boldsymbol{\beta}_\gamma$ marginalizing out σ^2, we integrate out σ^2, leading to

$$\boldsymbol{\beta}_\gamma|(\boldsymbol{Y}, \boldsymbol{\omega}, \boldsymbol{b}, a, d, \boldsymbol{\gamma}) \sim t_{2a}\left\{\hat{\boldsymbol{\beta}}_\gamma, \frac{d}{a}(\boldsymbol{X}_\gamma'\boldsymbol{X}_\gamma)^{-1}\right\}.$$

For random effects $\boldsymbol{b}$, letting $\boldsymbol{\Psi} = \left(\boldsymbol{\mathcal{V}}'\boldsymbol{\mathcal{V}} + \sigma^2\boldsymbol{\mathcal{G}}^{-1}\right)$, with analogous $\boldsymbol{\Psi}_i = \left(\boldsymbol{V}_i'\boldsymbol{V}_i + \sigma^2\boldsymbol{G}^{-1}\right)$, the conditional distribution is

$$\boldsymbol{b}|(\boldsymbol{Y}, \sigma^2, \boldsymbol{G}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\omega}) \sim N\left\{\boldsymbol{\Psi}^{-1}\boldsymbol{\mathcal{V}}'\left(\boldsymbol{Y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}_\gamma - \boldsymbol{A}'\boldsymbol{\omega}\right), \left(\sigma^{-2}\boldsymbol{\Psi}\right)^{-1}\right\},$$

as is common in Bayesian linear mixed models (Fearn, 1975; Harville & Zimmerman, 1996; Lange, Carlin, & Gelfand, 1992; Lindley & Smith, 1972). For the precision matrix of the random effects, the conditional distribution follows a Wishart distribution,

$$\boldsymbol{G}^{-1}|(\boldsymbol{Y}, \boldsymbol{C}, \rho, \sigma^2, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\omega}) \sim \mathcal{W}\left\{\left(\boldsymbol{b}\boldsymbol{b}' + \rho\boldsymbol{C}\right)^{-1}, N + \rho\right\}.$$

B Proofs of Propositions 1 and 2

We reiterate Propositions 1–2 from the main text and provide proofs for them below.

Proposition 1. *Under a standard EM algorithm, the parameters in the M-step for LMM-PROBE are not always estimable. That is, the maximizer of* $Q_{EM}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$*, where* $\boldsymbol{\eta} = (\boldsymbol{\beta}\ \boldsymbol{\omega}\ \boldsymbol{\tau})'$*, is not always unique.*

Proposition 2. *Let* $Q_{CM}^{M1}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ *and* $Q_{CM}^{M2}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ *denote the two M-step quantities in the PX-ECM algorithm for LMM-PROBE. Assuming no perfect collinearity between* $\boldsymbol{X}_k$ *and* $\boldsymbol{V}$ *for any* k*, the maximizers of* $Q_{CM}^{M1}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ *and* $Q_{CM}^{M2}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ *always exist and are unique.*

Proof of Proposition 1. Consider a standard M-step for the regression parameters in (1) without any additional non-sparse predictors. We add the expanded parameter $\boldsymbol{\tau}$ for r-dimensional random effects. For the small p non-sparse situation, Liu, Rubin, and Wu (1998) considered this model (see Section 4.1 therein) and were able to demonstrate superior convergence to the non-expanded parameter version. In this situation, the M-step maximizes

$$Q_{\text{EM}}(\boldsymbol{\eta} \mid \Theta^{(t-1)}) = -E_{\boldsymbol{\gamma}\boldsymbol{\beta}}\left[\{\boldsymbol{Y} - \boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\}'\{\boldsymbol{Y} - \boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\} \mid \Theta^{(t-1)}\right]$$

where $\boldsymbol{U}(\boldsymbol{b}) = (\boldsymbol{X}\ \boldsymbol{V}\ \boldsymbol{\mathcal{V}}'\boldsymbol{b})'$ and $\boldsymbol{\eta}(\boldsymbol{\gamma}) = (\boldsymbol{\gamma\beta}\ \boldsymbol{\omega}\ \boldsymbol{\tau})$ with $\boldsymbol{\eta} \equiv \boldsymbol{\eta}(\mathbf{1})$. For brevity, we drop the $(t-1)$ superscript on all expectations and, without loss of generality, assume $p_k > 0$ (predictor $\boldsymbol{X}_k$ can be taken out of the optimization when $p_k = 0$). Note that

$$E_{\boldsymbol{\gamma b}}\left\{\boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\right\} = \boldsymbol{U}(\tilde{\boldsymbol{b}})'\boldsymbol{\eta}(\boldsymbol{p}) \equiv \begin{bmatrix} (\boldsymbol{X} \odot \boldsymbol{P})' \\ \boldsymbol{V}' \\ \boldsymbol{\mathcal{V}}'\tilde{\boldsymbol{b}} \end{bmatrix}' \boldsymbol{\eta} = \tilde{\boldsymbol{U}}'\boldsymbol{\eta},$$

where $\odot$ denotes the Hadamard product and $\boldsymbol{P}$ is an $n \times p$ matrix in which each row is the vector $\boldsymbol{p}$, and

$$E_{\boldsymbol{\gamma b}}\left[\left\{\boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\right\}'\left\{\boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\right\}\right] = (\tilde{\boldsymbol{U}}'\boldsymbol{\eta})'\tilde{\boldsymbol{U}}'\boldsymbol{\eta} + \mathrm{Var}\left\{\boldsymbol{U}(\boldsymbol{b})'\boldsymbol{\eta}(\boldsymbol{\gamma})\right\}.$$

The variance term reduces to

$$\mathbf{1}'\begin{bmatrix} (\boldsymbol{X}^2) \odot \boldsymbol{P} \odot (\boldsymbol{J} - \boldsymbol{P}) \\ \mathbf{0}_r \\ \boldsymbol{\mathcal{V}}'(\sigma^{-2}\boldsymbol{\Psi})^{-1}\boldsymbol{\mathcal{V}} \end{bmatrix}\boldsymbol{\eta}^2 = \boldsymbol{\lambda}'\boldsymbol{\eta}^2$$

where $\boldsymbol{X}^2 = \boldsymbol{X} \odot \boldsymbol{X}$, $\boldsymbol{J}$ is a $(p \times p)$ matrix of ones, $\mathbf{0}_r$ is an r-dimensional vector of zeros, and $\boldsymbol{\lambda} = \left\{\lambda_1, \ldots, \lambda_p, \mathbf{0}_r', \boldsymbol{\mathcal{V}}'(\sigma^{-2}\boldsymbol{\Psi})^{-1}\boldsymbol{\mathcal{V}}\right\}'$ where $\lambda_k = p_k(1-p_k)\sum_i X_{ik}^2$ for $k \leq p$. This yields

$$Q_{\mathrm{EM}}(\boldsymbol{\eta} \mid \Theta^{(t-1)}) = -\left\{\left(\boldsymbol{Y} - \tilde{\boldsymbol{U}}'\boldsymbol{\eta}\right)'\left(\boldsymbol{Y} - \tilde{\boldsymbol{U}}'\boldsymbol{\eta}\right)\right\} - \left[\boldsymbol{\lambda}'\boldsymbol{\eta}^2\right],$$

resulting in maximizer

$$\hat{\boldsymbol{\eta}} = \left[\tilde{\boldsymbol{U}}'\tilde{\boldsymbol{U}} + \boldsymbol{\lambda}\boldsymbol{I}_{p+2r}\right]^{-1}\tilde{\boldsymbol{U}}'\boldsymbol{Y}.$$

As a result, $\hat{\boldsymbol{\eta}}$ is only estimable when $\sum_k I(p_k = 1) < M+1-r$; adding additional non-sparse predictors would lower this limit. Therefore, $Q_{\mathrm{EM}}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ cannot always be maximized, as the M-step is undefined when $\sum_k I(p_k = 1) \geq M + 1 - r$. □

Remark 1: A standard M-step for calculating $\hat{\boldsymbol{\eta}}$ requires the inverse of $\tilde{\boldsymbol{U}}'\tilde{\boldsymbol{U}} + \boldsymbol{\lambda}\boldsymbol{I}_{p+2r}$ a $(p + 2r) \times (p + 2r)$ matrix, which is lower-bounded by $\Omega(p^3)$ computational complexity. Note that this matrix is a function of the moments of $\boldsymbol{\gamma}$ and $\boldsymbol{b}$, which change at every iteration. Consequently, the inversion is required at every iteration, which compounds the computational complexity of the standard M-step. As demonstrated in the simulation studies in the main text, methods that require this type of inversion (e.g., PGEE) scale poorly with p and require computation times that are not feasible for our data analysis.

Remark 2: The benefit of the estimate of, say, β_k from the standard M-step is that it would adjust for the remaining predictors in the model. This points to the

motivation for including $\boldsymbol{\alpha}$, the expanded parameters for $\boldsymbol{\beta}$, in the proposed LMM-PROBE method. These expanded parameters adjust for the remaining signal, denoted by $\boldsymbol{W}_k$ in the main text, while estimating β_k. Without α_k, β_k would be estimated using $(\boldsymbol{Y}-\boldsymbol{W}_k)$ which is *corrected for* $\boldsymbol{W}_k$ instead of *adjusted for* $\boldsymbol{W}_k$. The differences in the estimate of β_k with correction versus adjustment can be large when $\boldsymbol{X}_k$ and $\boldsymbol{W}_k$ are related.

Proof of Proposition 2. In the PX-ECM algorithm for LMM-PROBE, the ℓth PX-CM partition maximizes the expected posterior of $\boldsymbol{\xi}_\ell$ (where $\boldsymbol{\eta}$ is a subset of $\boldsymbol{\xi}$, $\boldsymbol{\eta} \in \boldsymbol{\xi}$), which is a function of $\boldsymbol{X}_\ell$, $\boldsymbol{V}$, and the latent terms $\boldsymbol{U}_\ell$,

$$\hat{\boldsymbol{\xi}}_\ell^{(M1)} = \text{argmax}_{\boldsymbol{\xi}_\ell} E_{\boldsymbol{U}_\ell}\left\{ l(\boldsymbol{\xi}_\ell|\boldsymbol{Y}, \boldsymbol{U}_\ell, \boldsymbol{\Gamma}_\ell)|\Theta_{/\ell}^{(t-1)} \right\} \text{ for } \ell = 0, 1, \ldots, p.$$

With the parameter expansion, this reduces the first M-step (M1) to a straightforward calculation for each partition ℓ, akin to a $2(r+1)$-dimensional linear regression with maximizer

$$\hat{\boldsymbol{\xi}}_\ell^{(M1)} = \left\{ (\boldsymbol{Z}_\ell' \boldsymbol{Z}_\ell)^{(t-1)} \right\}^{-1} \boldsymbol{Z}_\ell^{(t-1)'} \boldsymbol{Y},$$

where $\boldsymbol{Z}_\ell$ is defined in Section 2.2 of the main text. This maximizer is fully identifiable under the PX-ECM framework as long as there is not perfect collinearity between $\boldsymbol{X}_k$ and $\boldsymbol{V}$ for all k. As a result, $Q_{\text{CM}}^{M1}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ can be maximized. The second M-step M2 updates $\hat{\boldsymbol{\xi}}_0^{(M2)}$, a subset of $\boldsymbol{\eta}$, conditional on the first M-step M1 and E-steps. This yields the same proof that the maximizer is identifiable and exists. Therefore, the maximizers of $Q_{\text{CM}}^{M1}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ and $Q_{\text{CM}}^{M2}(\boldsymbol{\eta} \mid \Theta^{(t-1)})$ both exist and are unique. □

C Convergence Assessment

This Section provides additional results and discussion with regards to convergence. Our implementation of LMM-PROBE is an 'all-at-once' optimization, where the parameter value updates at iteration t are performed accounting for updates from the previous iteration $t-1$ (analogous to the Jacobi least-squares optimization approach, Mascarenhas, 1995). In previous work, we have found that the 'all-at-once' optimization was less sensitive to the updating order and more efficiently maximized the likelihood than the 'one-at-a-time' approach (McLain, Zgodic, & Bondell, 2022). In the 'one-at-a-time' approach, parameter values are sequentially updated, with each update accounting for all previous updates within current iteration t (analogously to Gauss-Seidel least-squares optimization, Ma, Needell, & Ramdas, 2015). The 'one-at-a-time' approach fulfills the monotonicity and space-filling properties of EM and ECM algorithms (Meng & Rubin, 1993). Our 'all-at-once' implementation of LMM-PROBE is less prone to getting stuck in local maxima (in practice) compared to a 'one-at-a-time' approach but does not provide convergence guarantees.

Various studies have demonstrated that the PX-EM algorithm and some parameterizations of ECM have better global rates of convergence than a standard EM algorithm (Liu et al., 1998; Meng, 1994; Sexton & Swensen, 2000). However, for LMM-PROBE, the parameters in the complete data model of the standard EM algorithm are nonidentifiable when $p \gg n$ (Proposition 1), and thus the M-step is often undefined.

Therefore, we cannot compare the theoretical rate of convergence of a PX multi-cycle ECM as in LMM-PROBE to that of a standard EM. Instead, we assess empirical evidence of convergence. For eight simulation settings chosen to illustrate all simulation factors, Figure C.1 shows that the log-likelihood sharply increased in early iterations and flattened quickly in a small number of iterations, indicating a rapid rate of convergence. The empirical convergence assessment was similar for other simulation settings (omitted from figures).

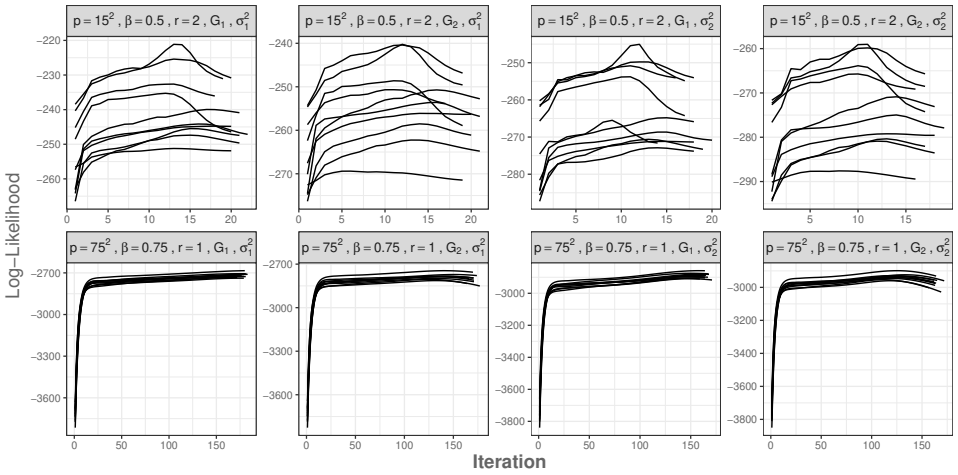

Fig. C.1 Log-Likelihood for LMM-PROBE over iterations, across various simulation settings, including r, σ^2, $\boldsymbol{G}$, and $\boldsymbol{\beta}$ values, when $p = (15^2, 75^2)$ and $\pi = 0.1$. Within each simulation scenario, lines represent simulation repetitions (10 repetitions).

To empirically assess the maximization of the conditional likelihood, we examined $-\frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}\sum_i \left[\boldsymbol{Y}_i - \left\{\boldsymbol{X}_i(\tilde{\alpha}_0\tilde{\boldsymbol{p}}\tilde{\boldsymbol{\beta}}) + \boldsymbol{V}_i\tilde{\boldsymbol{\omega}}_0 + \boldsymbol{V}_i\tilde{\boldsymbol{b}}_i\right\}\right]^2$ for all methods. Figure C.2 shows that LMM-PROBE maximized the conditional likelihood better than LASSO and PROBE in the settings displayed. Additionally, LMM-PROBE maximized the conditional likelihood better than LASSO+ in all settings except two, and better than LMM-LASSO when $\boldsymbol{\beta} = 0.5$ and variances were higher. When $\pi = 0.05$, LMM-PROBE maximized the conditional likelihood most effectively in all settings (figure omitted). Generally, methods for LMMs maximized the conditional likelihood better than LASSO and PROBE, which are not designed for LMMs.

D Connection with Sequentially Updated g-Prior

There is a connection between LMM-PROBE's β_k estimates after the mixing step (part of the E-step) and a prior we call the sequentially updated g-prior (Zellner, 1986). The first term of the MAP of the g-prior is equivalent to β_k under certain circumstances. We present the details of this connection here. To see the connection, note that if $\boldsymbol{Y} = \boldsymbol{\xi}\boldsymbol{Z} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ and $\boldsymbol{Z}$ as in the main text, then the prior for $\boldsymbol{\xi}$ is

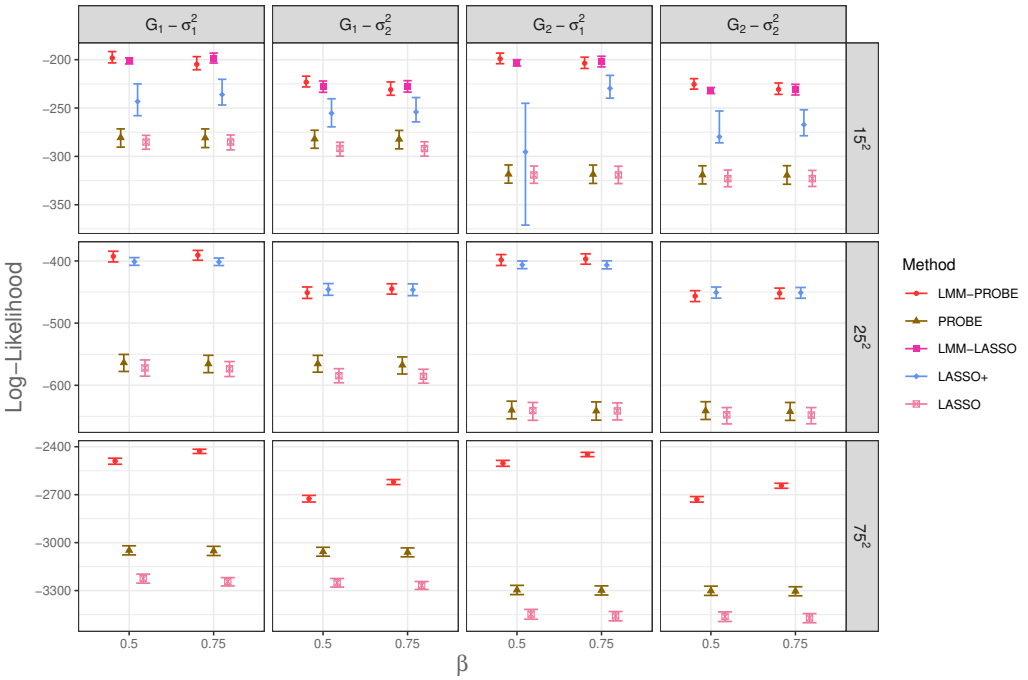

Fig. C.2 Log-Likelihood for LMM-PROBE and four comparison methods (excluding PGEE) across various simulation settings, including p, σ^2, $\boldsymbol{G}$, and $\boldsymbol{\beta}$ values, when $r = 2$ and $\pi = 0.1$. Vertical lines display the interquartile range of the Log-Likehood values across simulation iterations. Comparison methods LMM-LASSO and LASSO+ are methods for linear mixed models or repeated measures, while LASSO and PROBE are methods for linear models.

$\boldsymbol{\xi} \sim N\{\boldsymbol{\mu_0}, g\sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\}$. Then, the posterior distribution of $\boldsymbol{\xi}$ is $\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{g\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\hat{\boldsymbol{\xi}} + \frac{1}{g\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\mu_0}\right), \quad \text{and}$$
$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{g\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1},$$

where $\hat{\boldsymbol{\xi}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}$. This can be simplified to

$$\begin{aligned}\boldsymbol{\mu} &= \frac{g\sigma^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}}{1+g}\left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\hat{\boldsymbol{\xi}} + \frac{1}{g\sigma^2}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\mu_0}\right) \\ &= \frac{g}{1+g}\hat{\boldsymbol{\xi}} + \frac{1}{1+g}\boldsymbol{\mu_0}. \qquad \text{(D.1)}\end{aligned}$$

Let $\beta_k^{(t_{E2}-1)}$ be the estimate at the end of iteration $t-1$ and $\hat{\beta}_k^{(t_{M2})}$ be the first element of $\hat{\boldsymbol{\xi}}_k^{(t_{M2})} = \left(\hat{\beta}_k^{(t_{M2})}, \hat{\omega}_k^{(t_{M2})}, \hat{\boldsymbol{\phi}}_k^{(t_{M2})}\right)' = \left\{(\boldsymbol{Z}_k'\boldsymbol{Z}_k)^{(t_{E2}-1)}\right\}^{-1} \boldsymbol{Z}_k^{(t_{E2}-1)'}\boldsymbol{Y}$. Let $S_k^{2(t_{E2}-1)}$ and $\hat{S}_k^{2(t_{M2})}$ be defined as in Section 3 of the main text. Then, the mixing step (part of the LMM-PROBE E-step) is given by

$$\beta_k^{(t_{E2})} = (1-q^{(t_{E2})})\beta_k^{(t_{E2}-1)} + q^{(t_{E2})}\hat{\beta}_k^{(t_{M2})}, \text{ and} \tag{D.2}$$

$$S_k^{2(t_{E2})} = \left\{(1-q^{(t_{E2})})(S_k^{2(t_{E2}-1)})^{-1} + q^{(t_{E2})}(\hat{S}_k^{2(t_{M2})})^{-1}\right\}^{-1}, \tag{D.3}$$

where $\beta_k^{(0_{E2})} = 0$ and $S_k^{2(0_{E2})} = \infty$.

The estimate $\beta_k^{(t_{E2})}$ after the mixing step in (D.2) is equivalent to the first element of (D.1) if $q^{(t_{E2})} = \frac{g}{1+g}$, $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_k^{(t_{M2})}$, and $\boldsymbol{\mu_0} = (\beta_k^{(t_{E2}-1)}, 0, 0)'$. We note that the prior means of ω_k and ϕ_k do not impact the algorithm since they do not impact the estimate of β_k and the estimates of ω_k and ϕ_k are not used in any subsequent steps. Recalling that $q^{(t_{E1})} = (t+1)^{-1}$, the value of $\beta_k^{(t_{E2})}$ after mixing is equivalent to the MAP if the following prior is used

$$\boldsymbol{\xi}_k \sim N\left\{\begin{pmatrix}\beta_k^{(t_{E2}-1)}\\ 0\\ 0\end{pmatrix}, \frac{\sigma^2}{t}\left(\boldsymbol{Z}_k'\boldsymbol{Z}_k\right)^{-1}\right\},$$

All the elements of $\boldsymbol{\xi}_k$, i.e., $(\beta_k, \omega_k, \boldsymbol{\phi}_k)'$, have a proper prior in this view. However, the above g-prior will underestimate the posterior variance of β_k, which is required for the E-step. That is, posterior variance from the g-prior

$$\frac{\sigma^2}{t}\left(\boldsymbol{Z}_k'\boldsymbol{Z}_k\right)^{-1} \to 0 \text{ as } t \to \infty.$$

This arises because the g-prior does not account for the uncertainty in $\beta_k^{(t_{E2}-1)}$, rather it treats $\beta_k^{(t_{E2}-1)}$ as fixed. As a result, in the LMM-PROBE framework, we present the estimation procedure using a non-informative prior for $\boldsymbol{\xi}_k$ and a mixing step, since the variance in (D.3) incorporates the uncertainty in $\beta_k^{(t_{E2}-1)}$.

E Additional Simulation Results

This Section provides additional simulation results. Figure E.3 corresponds to Figure 1 in Section 4 of the main text, with the additional method of penalized generalized estimating equations (PGEE) for high-dimensional longitudinal data. The Mean Squared Predictive Error (MSPE) was the highest for this additional method. Figure E.4 also mirrors Figure 1 from Section 4 in the main text, but shows MSPEs for the simulation settings with one random effect ($r = 1$). Figure E.5 shows Median Absolute Deviations (MADs) for simulation settings when $r = 1$ and $\pi = 0.1$. In all settings

except one, MADs were lowest for LMM-PROBE. Figures E.6 and E.7 show sensitivity, specificity, and the Matthews Correlation Coefficient (MCC, Matthews, 1975) when $p = 25^2$ and $p = 75^2$, respectively. Figure E.6 shows that when the effect size of signals was higher ($\boldsymbol{\beta} = 0.75$), LMM-PROBE had a similar sensitivity as PROBE, and outperformed other methods. LMM-PROBE had the highest specificity for all settings. In Figure E.7, LMM-PROBE had higher sensitivity when $\boldsymbol{\beta} = 0.75$ as well as high sensitivity in all settings except one.

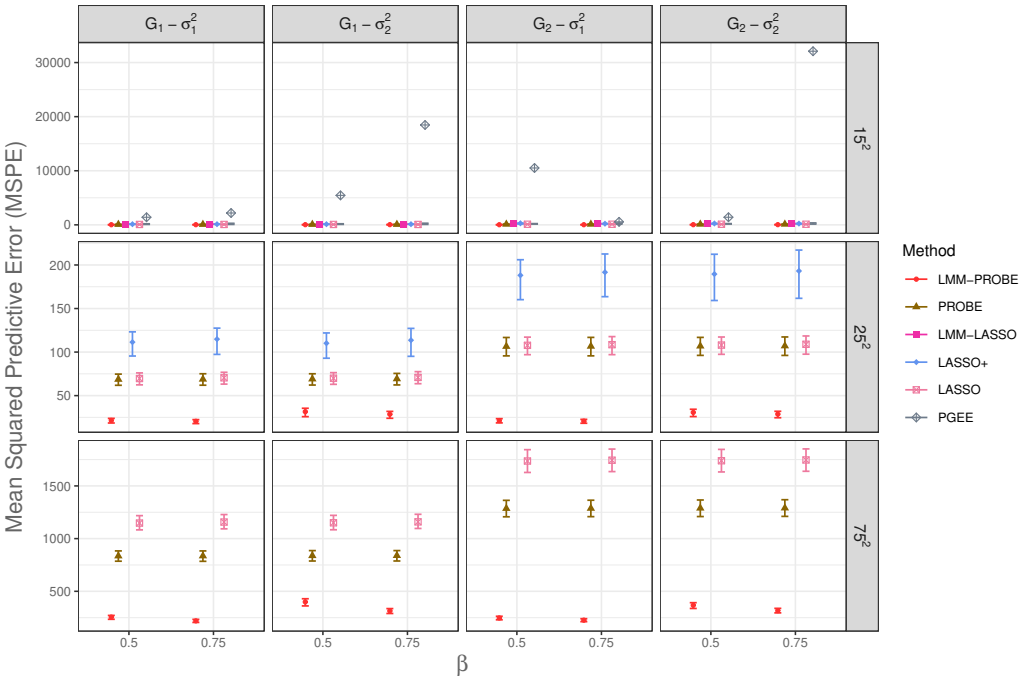

Fig. E.3 Mean Squared Predictive Errors (MSPE) for LMM-PROBE and five comparison methods (including PGEE) across various simulation settings, including p, σ^2, $\boldsymbol{G}$, and $\boldsymbol{\beta}$ values, when $r = 2$ and $\pi = 0.1$. The MSPEs are based on both fixed and random effects. Vertical lines display the interquartile range of the MSPEs. Comparison methods LMM-LASSO, LASSO+, and PGEE are methods for linear mixed models or repeated measures, while LASSO and PROBE are methods for linear models.

F Additional SLE Data Analysis Results

In this Section, we examine additional results from the data analysis on pediatric systemic lupus erythematosus (SLE). For the SLE analysis, Figure F.8 shows MSPEs and MADs for LMM-PROBE, PROBE, and LASSO. Note that LASSO and PROBE are not suitable for linear mixed modeling. Here, we show that by ignoring the dependence structure in the data, LASSO and PROBE methods do not accurately reflect the true contributions of the random effects to the response variable, leading to less

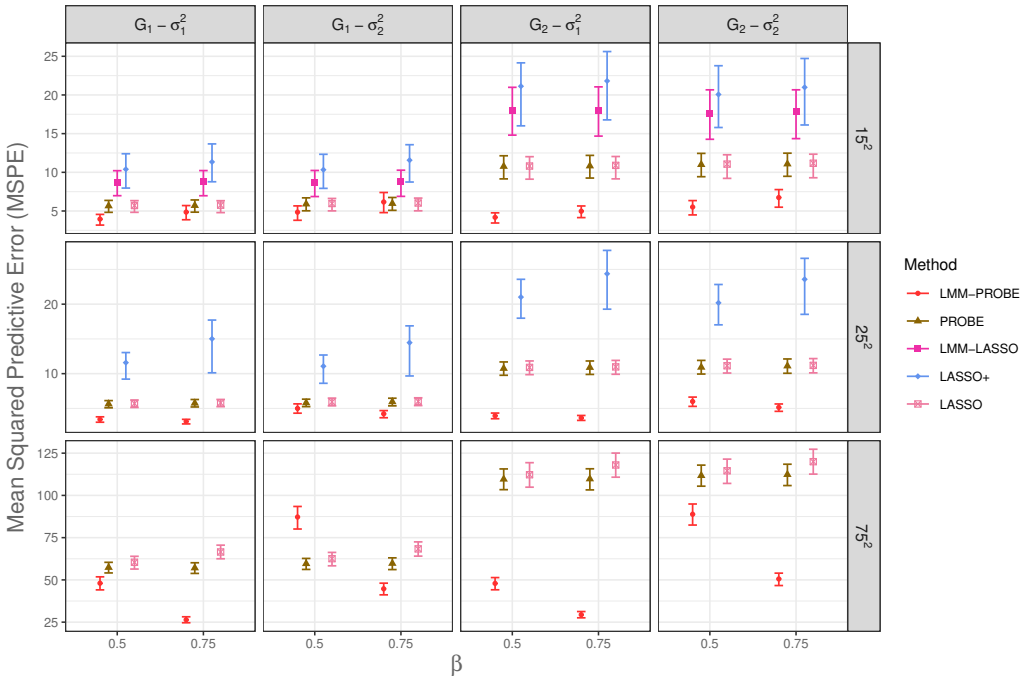

Fig. E.4 Mean Squared Predictive Errors (MSPE) for LMM-PROBE and four comparison methods across various simulation settings, including p, σ^2, $\boldsymbol{G}$, and $\boldsymbol{\beta}$ values, when $r = 1$ and $\pi = 0.1$. The MSPEs are based on both fixed and random effects. Vertical lines display the interquartile range of the MSPEs. Comparison methods LMM-LASSO and LASSO+ are methods for linear mixed models, while LASSO and PROBE are methods for linear models.

parsimonious models with lower out-of-sample prediction accuracy. For MSPEs, LMM-PROBE performed best while LASSO ranked last, and the range of MSPE values across the CV folds was wider for PROBE and LASSO compared to LMM-PROBE. Trends in MADs results were similar to those of MSPEs. However, PROBE had the lowest MAD.

Figure F.9 shows estimates for the within- and between-cluster variances, $\tilde{\sigma}^2$ and $\boldsymbol{G}$. For PROBE and LASSO, $\tilde{\sigma}^2$ is the model's residual variance estimate since these methods do not delineate within-unit and between-unit variation. LMM-PROBE had the lowest estimates of $\tilde{\sigma}^2$ across CV folds (on average 0.1). The average random effect variance $\tilde{\boldsymbol{G}}$ estimate was approximately 0.02 giving an average ICC of 0.15. Overall, LMM-PROBE captured both the residual and random effect variances and provided a better MSPE than PROBE and LASSO, which do not estimate between-cluster variance.

We examined the average number of predictors selected by each method. For LASSO, a predictor was selected if $\beta_k \neq 0$ while for PROBE and LMM-PROBE, a predictor was selected if $\tilde{p}_k > 0.5$. Figure F.10 shows that LMM-PROBE, PROBE, and LASSO selected 5, 9, and 167 predictors on average, with LMM-PROBE selecting the fewest predictors across CV folds. The methods for traditional high-dimensional

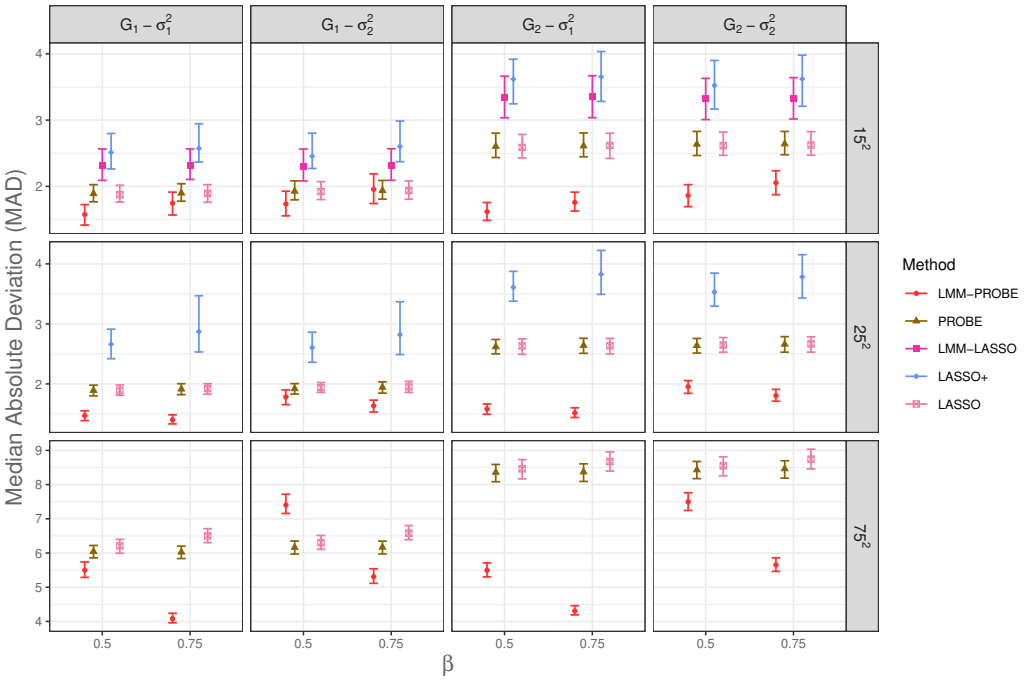

Fig. E.5 Median Absolute Deviations (MAD) for LMM-PROBE and four comparison methods across various simulation settings, including p, σ^2, $\boldsymbol{G}$, and $\boldsymbol{\beta}$ values, when $r = 1$ and $\pi = 0.1$. The MADs are based on both fixed and random effects. Vertical lines display the interquartile range of the MADs. Comparison methods LMM-LASSO and LASSO+ are methods for linear mixed models, while LASSO and PROBE are methods for linear models.

linear regression (LASSO, PROBE) required around nine and 48 seconds per fold, respectively, while LMM-PROBE required 69 seconds on average.

We examined residuals for each CV fold. Figure F.11 shows conditional residuals (Nobre & Singer, 2007) plotted against observations for each CV fold. These residuals show that the homogeneous variance assumption is fulfilled. Quantile-Quantile (Q-Q) plots in Figure F.12 show that the normality assumption is not violated. Figure F.13 shows the predicted random intercepts plotted against subjects, where no subject had an outlying predicted random intercept. Finally, the Q-Q plots in Figure F.14 show that the random effect normality assumption is not violated either.

G Data Analysis on Riboflavin Dataset

In this Section, we showcase the performance and characteristics of the LMM-PROBE method on the riboflavin (vitamin B_2) dataset (Buhlmann, Kalisch, & Meier, 2014), a popular high-throughput genomic dataset about riboflavin production using a genetically modified bacterium. Riboflavin is grown in recombinant *Bacillus subtilis*, with the goal of maintaining the production rate of the riboflavin over time. Riboflavin grows across multiple generations of the bacterium for a different duration, leading to a longitudinal design with repeated observations. The dataset contains a log-transformed

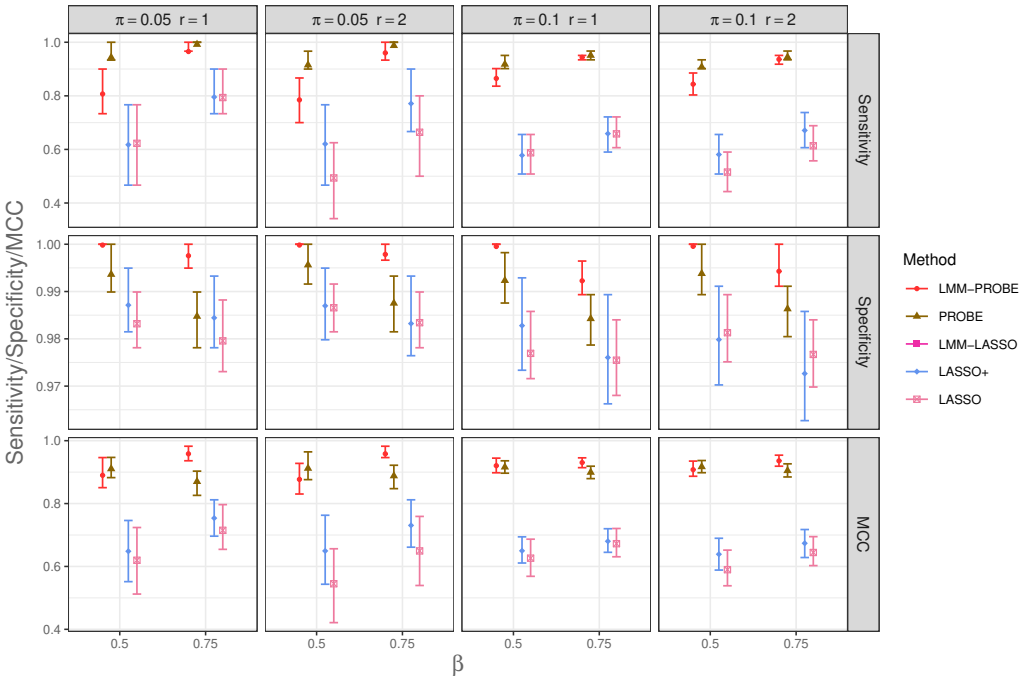

Fig. E.6 Sensitivity, Specificity, and the Matthews Correlation Coefficient (MCC) for LMM-PROBE and four comparison methods across various simulation settings, including π, r, and $\boldsymbol{\beta}$ values, when $\sigma^2 = \sigma_1^2$, $\boldsymbol{G} = \boldsymbol{G}_1$, and $p = 25^2$. Vertical lines display the interquartile range of the sensitivity, specificity, and MCC. Comparison methods LMM-LASSO and LASSO+ are methods for linear mixed models, while LASSO and PROBE are methods for linear models.

riboflavin production rate, which we use as the outcome, as well as 4,088 predictor variables, each consisting of gene expression levels on the log scale. There are 28 clusters in the dataset, measured over time, with the number of observations per cluster ranging between 2 and 6. In total, there are 111 observations. The original aim of the research that generated this dataset was to identify the genes that impacted riboflavin production. To analyze this data, we used a linear mixed-effects model. We considered all 4,088 gene expression predictors as sparse fixed effects and considered an intercept as well as a time predictor as both non-sparse fixed and random effects.

To evaluate the performance of LMM-PROBE, we used MSPEs resulting from five-fold CV. In the CV, we balanced clusters across the folds, meaning that a cluster's observations were all in the same fold. At a given iteration of the CV, 80% of the clusters were in training folds, and 20% were in the validation-test fold. The validation-test fold was further split into validation (with earlier *time* values for each cluster in the fold, i.e., 1–2) and testing (with subsequent *time* values, i.e., 3–6) subfolds. The testing subfold had two observations from each cluster with four or more measurements and one observation otherwise. We used the validation subfold to obtain the predicted random effects and the testing subfold to calculate MSPEs. To generate random effect predictions for the validation subfold, we used the MAP estimate of LMM-PROBE.

Fig. E.7 Sensitivity, Specificity, and the Matthews Correlation Coefficient (MCC) for LMM-PROBE and two comparison methods across various simulation settings, including π, r, and $\boldsymbol{\beta}$ values, when $\sigma^2 = \sigma_1^2$, $\boldsymbol{G} = \boldsymbol{G}_1$, and $p = 75^2$. Vertical lines display the interquartile range of the sensitivity, specificity, and MCC. Comparison methods LASSO and PROBE are methods for linear models.

To tune the penalty parameter in LASSO, we performed an additional five-fold CV using the training set.

We calculated two types of MSPEs. The first type of MSPE is the difference between the true outcome value and prediction obtained based on both the fixed and random effect estimates. Specifically, using estimates $\tilde{\boldsymbol{G}}$ and $\tilde{\sigma}^2$ from the training data, we obtained predicted random effects $\tilde{\boldsymbol{b}}_i$ using the validation data and finally obtained complete predictions for future time points using the testing data. The second type of MSPE is the difference between the true outcome value and predictions obtained only based on the fixed effect estimates. Figure G.15 shows the two types of MSPEs as well as MADs. For MSPEs based on both fixed and random effect estimates, LMM-PROBE had the lowest MSPEs compared to PROBE and LASSO. When examining MSPEs based on fixed effect estimates only, LMM-PROBE performed best as well. Generally, the range of MSPE values across the CV folds was wider for LASSO and PROBE, and narrower for LMM-PROBE. Trends in MADs results were overwhelmingly similar to those of MSPEs.

Figure G.16 shows estimates for within-sample variation, $\tilde{\sigma}^2$. For PROBE and LASSO, $\tilde{\sigma}^2$ is the model residual error estimates since these methods do not delineate between within-unit and between-unit variation. LMM-PROBE had a mid-range estimate for $\tilde{\sigma}^2$ with the lowest variability across CV fold. LASSO had the lowest average

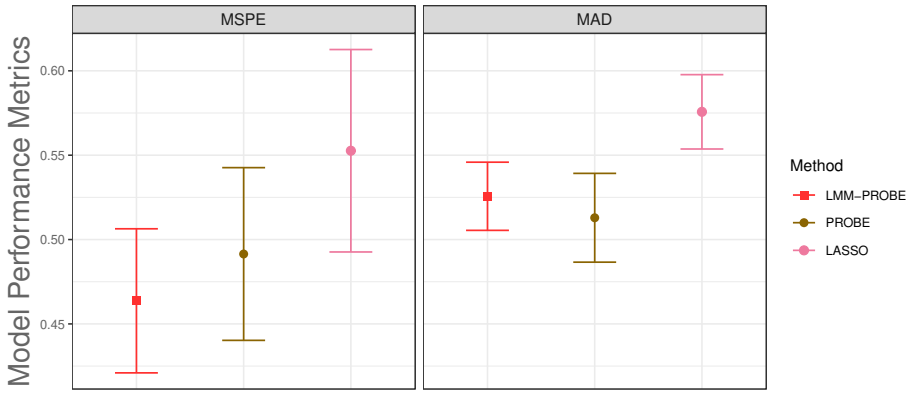

Fig. F.8 Mean Squared Predictive Errors (MSPE) and Median Absolute Deviations (MAD) for LMM-PROBE and two comparison methods. Vertical lines represent ± the standard error of MSPE or MAD, divided by $\sqrt{5}$, based on the number of Cross-Validation folds.

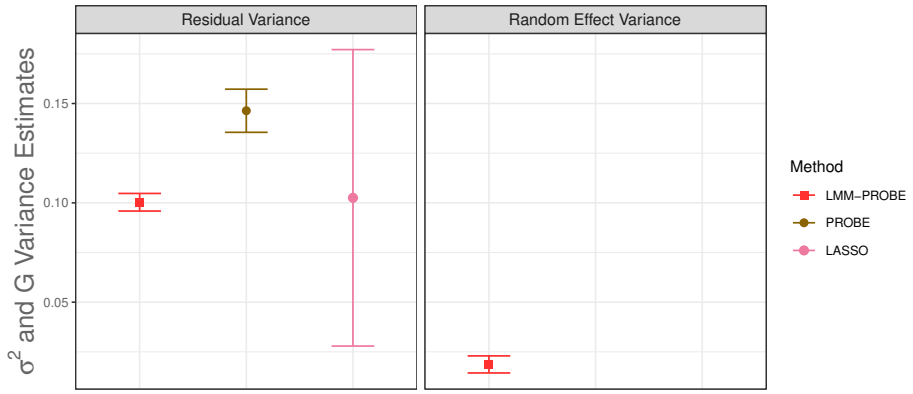

Fig. F.9 Average residual ($\tilde{\sigma}^2$) and random effect ($\boldsymbol{G}$) variance estimates for LMM-PROBE and two comparison methods. Vertical lines represent ± the standard error of $\tilde{\sigma}^2$ or $\boldsymbol{G}$, divided by $\sqrt{5}$, based on the number of Cross-Validation folds.

$\tilde{\sigma}^2$ estimate but varied markedly across CV folds. Figure G.17 shows the average number of predictors selected by LMM-PROBE (1), PROBE (1), and LASSO (33), with LMM-PROBE selecting the fewest predictors across CV folds. Prior to choosing the final model described above with the LMM-PROBE, LASSO, and PROBE approaches, we examined additional methods (LASSO+, LASSO+EN, and BRMS) but did not continue with them due to computation time. Figure G.18 shows the average computation time in seconds for one CV fold of an intermediate model in our data analysis. The methods for traditional high-dimensional linear regression (LASSO, PROBE) required around one second per fold. The remaining methods required between 200 and 6,000 seconds, on average, per fold. Overall, the total computation time for the intermediate model was 1,026 minutes for BRMS, 202 minutes for LASSO+EN, 51 minutes for

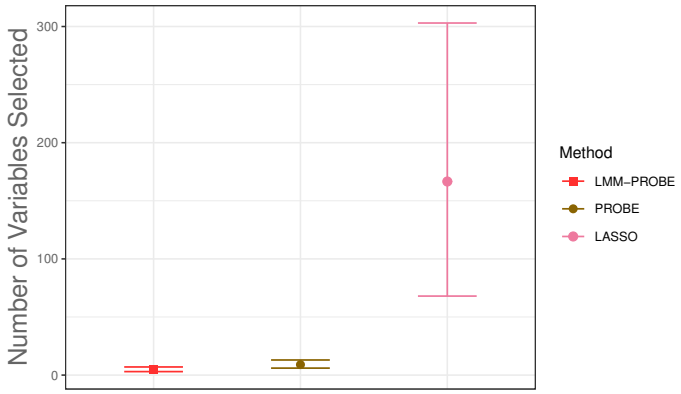

Fig. F.10 Average number of predictors selected for LMM-PROBE and two comparison methods. Vertical lines display the range across the Cross-Validation folds.

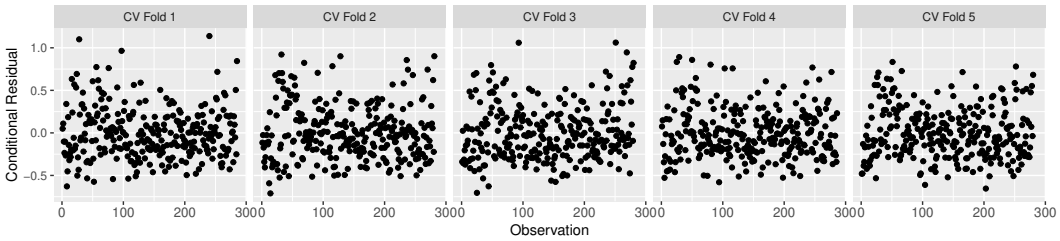

Fig. F.11 Conditional residuals plotted against observations for each Cross-Validation fold.

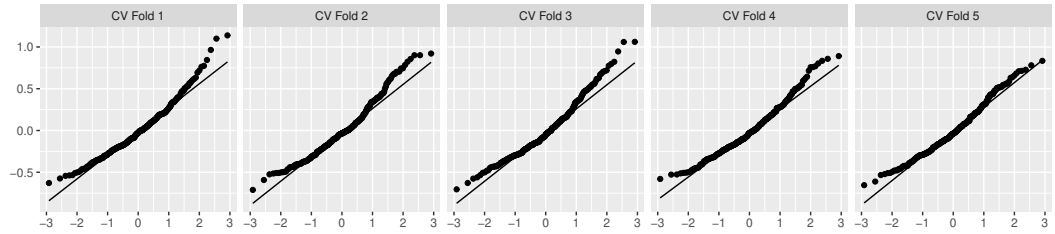

Fig. F.12 Quantile-Quantile plots of conditional residuals for each Cross-Validation fold.

LASSO+, 2 minutes for LMM-PROBE, 31 seconds for PROBE, and 18 seconds for LASSO.

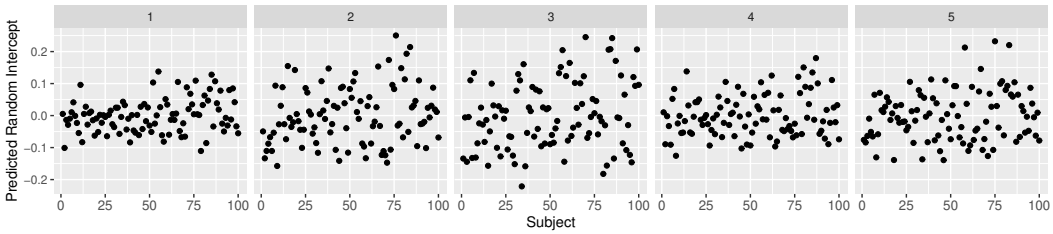

Fig. F.13 Predicted random intercepts plotted against subjects for each Cross-Validation fold.

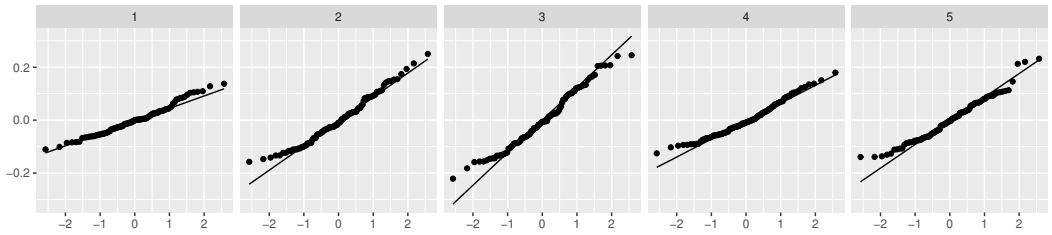

Fig. F.14 Quantile-Quantile plots of predicted random intercepts for each Cross-Validation fold.

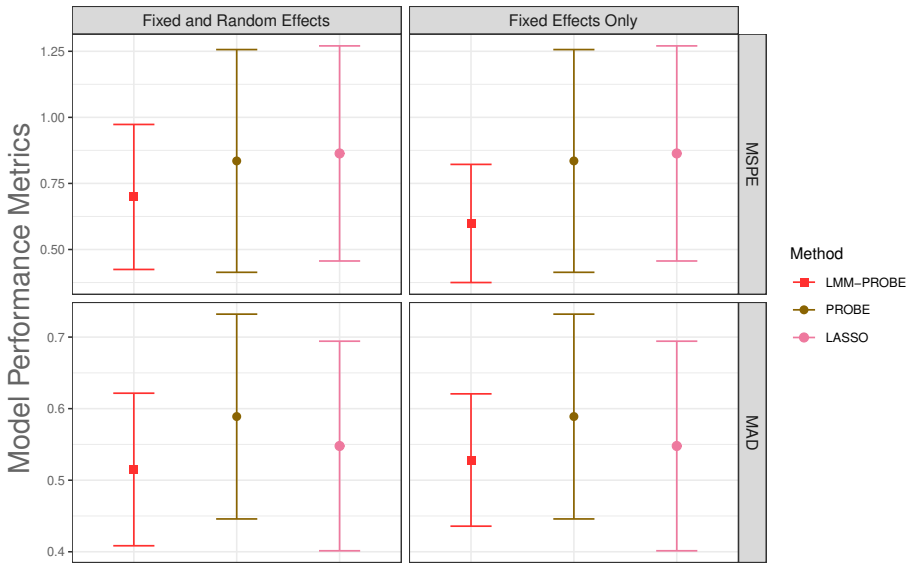

Fig. G.15 Mean Squared Predictive Errors (MSPE) and Median Absolute Deviations (MAD) for LMM-PROBE and two comparison methods, based on both random and fixed effects and fixed effects only. Vertical lines represent $\pm$ the standard error of MSPE or MAD, divided by $\sqrt{5}$, based on the number of Cross-Validation folds.

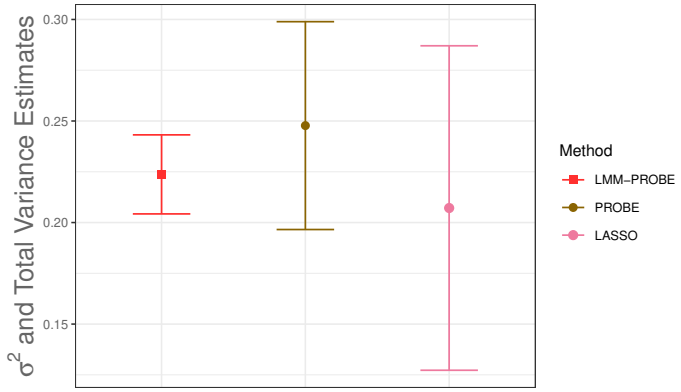

Fig. G.16 Average residual ($\tilde{\sigma}^2$) variance estimates for LMM-PROBE and two comparison methods. Vertical lines represent $\pm$ the standard error of $\tilde{\sigma}^2$ divided by $\sqrt{5}$, based on the number of Cross-Validation folds.

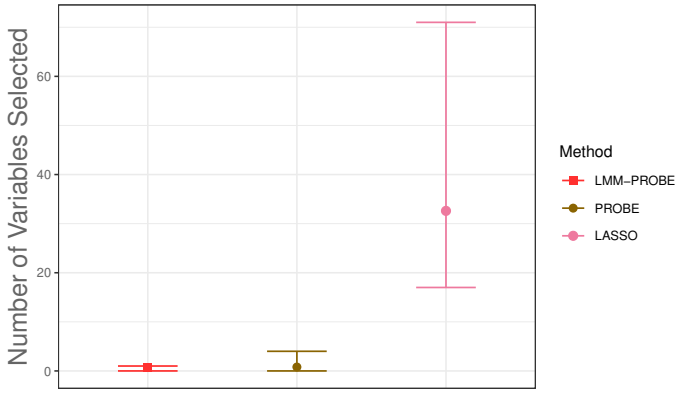

Fig. G.17 Average number of predictors selected for LMM-PROBE and two comparison methods. Vertical lines display the range across the Cross-Validation folds.

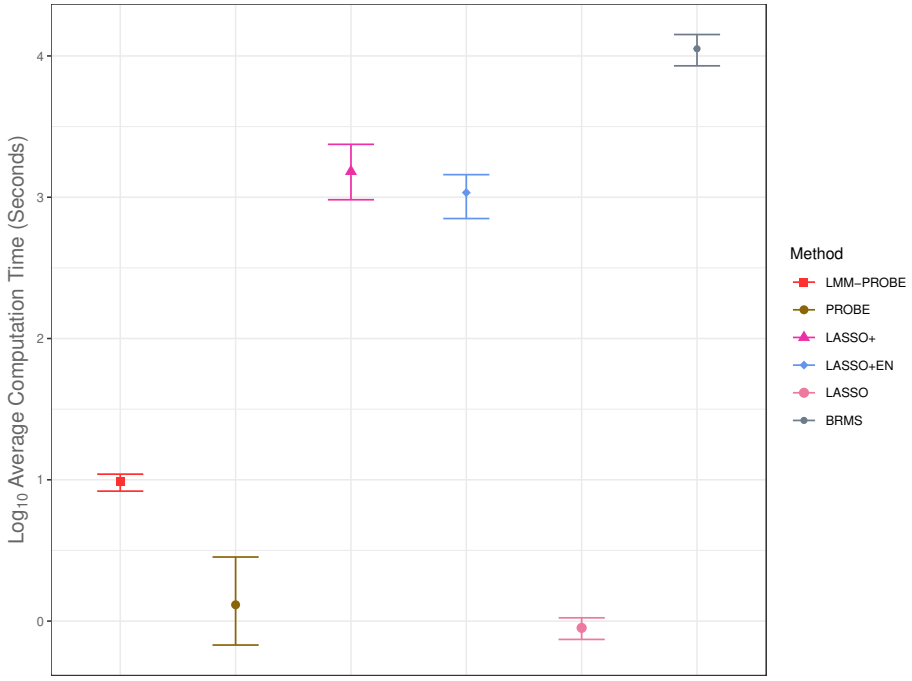

Fig. G.18 Average computation time per Cross-Validation fold (in seconds) for LMM-PROBE and five comparison methods, on the $\log_{10}$ scale.

References

Buhlmann, P., Kalisch, M., Meier, L. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, *1*, 255–278,

Fearn, T. (1975). A bayesian approach to growth curves. *Biometrika*, *62*(1), 89–100,

Harville, D.A., & Zimmerman, A.G. (1996). The posterior distribution of the fixed and random effects in a mixed-effects linear model. *Journal of Statistical Computation and Simulation*, *54*, 211–229,

Lange, N., Carlin, B.P., Gelfand, A.E. (1992). Hierarchical bayes models for the progression of hiv infection using longitudinal cd4 t-cell numbers. *Journal of the American Statistical Association*, *87*(419), 615–626,

Lindley, D., & Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(1), 1–18,

Liu, C., Rubin, D.B., Wu, Y.N. (1998, 12). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85*(4), 755-770, https://doi.org/10.1093/biomet/85.4.755 Retrieved from https://doi.org/10.1093/biomet/85.4.755

Ma, A., Needell, D., Ramdas, A. (2015). Convergence properties of the randomized extended gauss–seidel and kaczmarz methods. *SIAM Journal on Matrix Analysis and Applications*, *36*(4), 1590–1604,

Mascarenhas, W.F. (1995). On the convergence of the jacobi method for arbitrary orderings. *SIAM journal on matrix analysis and applications*, *16*(4), 1197–1209,

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451,

McLain, A.C., Zgodic, A., Bondell, H. (2022). *Sparse high-dimensional linear regression with a partitioned empirical bayes ecm algorithm.* https://arxiv.org/abs/2209.08139. arXiv. Retrieved from https://arxiv.org/abs/2209.08139

Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.*, *22*(1), 326–339, https://doi.org/10.1214/aos/1176325371 Retrieved from https://doi.org/10.1214/aos/1176325371

Meng, X.L., & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278, https://doi.org/10.1093/biomet/80.2.267

Nobre, J., & Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, *49*(6), 863–875,

Sexton, J., & Swensen, A.R. (2000). Ecm algorithms that converge at the rate of em. *Biometrika*, *87*(3), 651–662,

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, *6*, 233–243,