Supplementary Material to Hui et al., Joint Selection in Mixed Models using Regularized PQL

1 Proof of the estimation consistency in Theorem 1

To begin, we note that as discussed in the main text, conditions (C1)-(C4) and the independence of the clusters are sufficient to ensure the existence and $n^{1/2}$-consistency of the maximum likelihood estimators, $\tilde{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0 = O_p(n^{-1/2})$; see see also Lemmas 1 and 2 in Nie (2007).

For clarity in the development below, we shall explicitly write the regularized PQL as a function of $\boldsymbol{\beta}, \boldsymbol{\Gamma}$, and $\boldsymbol{b}$, such that $\ell_p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{b}) = \ell_{\text{PQL}}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{b}) - \lambda \sum_{k=1}^{p_f} v_k|\beta_k| - \lambda \sum_{l=1}^{p_r} w_l \|\boldsymbol{b}_{\bullet l}\|$. Let $\alpha_m = m_1^{-1/2}$, where we recall that $m_1 = O(m_{i'})$ for $i' = 2, \ldots, n$, and $\boldsymbol{b}_0 = (\boldsymbol{b}_{01}^T, \ldots, \boldsymbol{b}_{0n}^T)^T$. Define a vector $\boldsymbol{u} = (\boldsymbol{u}_1^T, \boldsymbol{u}_2^T)^T$ where $\boldsymbol{u}_1 = (u_1, \ldots, u_{p_f})$, $\boldsymbol{u}_2 = (\boldsymbol{u}_{21}^T, \ldots, \boldsymbol{u}_{2n}^T)^T$, $\boldsymbol{u}_{2i}$ is of length p_r, and $\|\boldsymbol{u}\|_\infty$ is bounded above by a large constant. Then consider the difference

$$
\begin{aligned}
\Delta(\boldsymbol{u}, \boldsymbol{\Gamma}) &= \frac{1}{n}\left\{\ell_p(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}_1, \boldsymbol{\Gamma}, \boldsymbol{b}_0 + \alpha_m \boldsymbol{u}_2) - \ell_p(\boldsymbol{\beta}_0, \boldsymbol{\Gamma}, \boldsymbol{b}_0)\right\} \\
&\leq \frac{1}{n}\left\{\ell_{\text{PQL}}(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}_1, \boldsymbol{\Gamma}, \boldsymbol{b}_0 + \alpha_m \boldsymbol{u}_2) - \ell_{\text{PQL}}(\boldsymbol{\beta}_0, \boldsymbol{\Gamma}, \boldsymbol{b}_0)\right\} \\
&\quad - \frac{1}{n}\left\{\lambda \sum_{k=2}^{p_{0f}} v_k \left(|\beta_{0k} + \alpha_m u_{1k}| - |\beta_{0k}|\right) + \lambda \sum_{l=1}^{p_{0r}} w_l \left(\|\boldsymbol{b}_{0\bullet l} + \alpha_m \boldsymbol{u}_{2l}\| - \|\boldsymbol{b}_{0\bullet l}\|\right)\right\} \\
&\triangleq T_1 - T_2,
\end{aligned}
$$

where the inequality in the second line follows since in the penalties, we sum only to p_{0f} and p_{0r} for the fixed and random effects respectively. From equation (1) in the main text, term T_1 may be decomposed as

$$T_1 = \frac{1}{n}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(y_{ij}|\boldsymbol{\beta}_0+\alpha_m\boldsymbol{u}_1,\boldsymbol{b}_{0i}+\alpha_m\boldsymbol{u}_{2i}) - \log f(y_{ij}|\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\right\}$$
$$-\frac{1}{2n}\sum_{i=1}^{n}\left\{(\boldsymbol{b}_{0i}+\alpha_m\boldsymbol{u}_{2i})^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^-(\boldsymbol{b}_{0i}+\alpha_m\boldsymbol{u}_{2i}) - \boldsymbol{b}_{0i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^-\boldsymbol{b}_{0i}\right\}$$
$$\triangleq U_1 - U_2.$$

We seek to prove that, for any $\boldsymbol{\Gamma}$ and all $m_i \to \infty$ and $n \to \infty$, U_1 is negative and dominates terms T_2 and U_2. First, let $\ell_1(\boldsymbol{\beta},\boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i)$. Then by a Taylor expansion, we have for U_1,

$$U_1 = \alpha_m\boldsymbol{u}^T\left\{\frac{1}{n}\nabla\ell_1(\boldsymbol{\beta}_0,\boldsymbol{b}_0)\right\} - \frac{\alpha_m^2}{2}\boldsymbol{u}^T\left\{-\frac{1}{n}\nabla^2\ell_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}})\right\}\boldsymbol{u}$$
$$\triangleq V_1 + V_2,$$

where $(\bar{\boldsymbol{\beta}}^T,\bar{\boldsymbol{b}}^T)^T$ lies on the line segment joining $(\boldsymbol{\beta}_0^T,\boldsymbol{b}_0^T)^T$ and $(\boldsymbol{\beta}_0^T+\alpha_m\boldsymbol{u}_1^T,\boldsymbol{b}_0^T+\alpha_m\boldsymbol{u}_2^T)^T$. Writing $\boldsymbol{\mu}_i(\boldsymbol{\beta},\boldsymbol{b}_i) = \{a'(\eta_{i1}),\dots,a'(\eta_{in})\}^T$, we have

$$\nabla\ell_1(\boldsymbol{\beta}_0,\boldsymbol{b}_0) = \left[\sum_{i=1}^{n}\boldsymbol{X}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\},\boldsymbol{Z}_1^T\{\boldsymbol{y}_1-\boldsymbol{\mu}_1(\boldsymbol{\beta}_0,\boldsymbol{b}_{01})\},\dots,\boldsymbol{Z}_n^T\{\boldsymbol{y}_n-\boldsymbol{\mu}_n(\boldsymbol{\beta}_0,\boldsymbol{b}_{0n})\}\right]. \quad (1)$$

Let $\mathrm{E}_{\boldsymbol{y}}(\cdot)$ and $\mathrm{Var}_{\boldsymbol{y}}(\cdot)$ denote the expectation and variance operators with respect to $\boldsymbol{y}$, respectively. Then for all $i = 1,\dots,n$, we have by the conditional independence of y_{ij} given $\boldsymbol{b}_{0i}$,

$$\mathrm{E}_{\boldsymbol{y}_i,\boldsymbol{b}_i}\left[\boldsymbol{X}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\}\right] = \boldsymbol{X}_i^T\mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{E}_{\boldsymbol{y}_i|\boldsymbol{b}_i}\left\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\right\}\right] = \boldsymbol{0},$$

$$\mathrm{Var}_{\boldsymbol{y}_i,\boldsymbol{b}_i}\left[\boldsymbol{X}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\}\right] = \boldsymbol{X}_i^T\mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{Var}_{\boldsymbol{y}_i|\boldsymbol{b}_i}\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\}\right]\boldsymbol{X}_i$$
$$= \boldsymbol{X}_i^T\mathrm{E}_{\boldsymbol{b}_i}\left\{\boldsymbol{W}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\right\}\boldsymbol{X}_i,$$

where $\boldsymbol{W}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})$ is a $m_i\times m_i$ diagonal matrix with entries $a''(\eta_{0ij})/\phi_0$, and $\eta_{0ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_0+\boldsymbol{z}_{ij}^T\boldsymbol{b}_{0i}$. Under Conditions (C1)-(C2) and (C4), applying Markov's inequality, it is straightforward to show that $\sum_{i=1}^n \boldsymbol{X}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\} = O_p(N^{1/2})$, so $n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^T[\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})] = O_p(N^{1/2}n^{-1})$. A similar development can be used to show that $\boldsymbol{Z}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\} = O_p(m_i^{1/2})$, and hence $n^{-1}\boldsymbol{Z}_i^T\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_0,\boldsymbol{b}_{0i})\} = O_p(m_i^{1/2}n^{-1})$. Combining these results and using the Cauchy-Schwarz inequality, we have $|V_1| = \left|\alpha_m\boldsymbol{u}^T\left\{n^{-1}\nabla\ell_1(\boldsymbol{\beta}_0,\boldsymbol{b}_0)\right\}\right| \leq \alpha_m\|\boldsymbol{u}\|\times\|n^{-1}\nabla\ell_1(\boldsymbol{\beta}_0,\boldsymbol{b}_0)\| = O_p(\{N/(nm_1)\}^{1/2})$, where we have the used the fact that $N=\sum_{i=1}^n m_i = O(nm_n)$. Turning to V_2, we see that

$$-\frac{1}{n}\nabla^2\ell_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^T\boldsymbol{W}_i(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_i)\boldsymbol{X}_i & \frac{1}{n}\boldsymbol{X}_1^T\boldsymbol{W}_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_1)\boldsymbol{Z}_1 & \cdots & \frac{1}{n}\boldsymbol{X}_n^T\boldsymbol{W}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_1)\boldsymbol{Z}_n \\ \frac{1}{n}\boldsymbol{Z}_1^T\boldsymbol{W}_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_1)\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{Z}_1^T\boldsymbol{W}_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_1)\boldsymbol{Z}_1 & & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\boldsymbol{Z}_n^T\boldsymbol{W}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_n)\boldsymbol{X}_n & \boldsymbol{0} & & \frac{1}{n}\boldsymbol{Z}_n^T\boldsymbol{W}_n(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}}_n)\boldsymbol{Z}_n \end{pmatrix}. \tag{2}$$

By condition (C1), $\alpha_m^2\boldsymbol{u}^T\left\{-n^{-1}\nabla^2\ell_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}})\right\}\boldsymbol{u} \geq \alpha_m^2 c_0\phi_0^{-1}\boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u}$ where

$$\boldsymbol{M} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^T\boldsymbol{X}_i & \frac{1}{n}\boldsymbol{X}_1^T\boldsymbol{Z}_1 & \cdots & \frac{1}{n}\boldsymbol{X}_n^T\boldsymbol{Z}_n \\ \frac{1}{n}\boldsymbol{Z}_1^T\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{Z}_1^T\boldsymbol{Z}_1 & & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\boldsymbol{Z}_n^T\boldsymbol{X}_n & \boldsymbol{0} & & \frac{1}{n}\boldsymbol{Z}_n^T\boldsymbol{Z}_n \end{pmatrix}.$$

Next, by conditions (C2) and (C4), we have the following order results for all $i = 1, \dots, n$: $n^{-1}\boldsymbol{Z}_i^T\boldsymbol{X}_i = O_p(m_i n^{-1})$, $n^{-1}\boldsymbol{Z}_i^T\boldsymbol{Z}_i = O_p(m_i n^{-1})$, and $n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i = O_p(Nn^{-1})$. It follows that

$$V_2 = -\frac{\alpha_m^2}{2}\boldsymbol{u}^T\left\{-\frac{1}{n}\nabla^2\ell_1(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}})\right\}\boldsymbol{u} \le -\frac{\alpha_m^2}{2}\frac{c_0}{\phi_0}\boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u} \le -1\times O_p\left(\frac{N}{nm_1}\right) = -1\times O_p\left(\frac{N}{m_1}\right).$$

Hence for n large enough, term V_2 asymptotically dominates V_1 and hence $U_1 = V_1 + V_2$ is negative.

It remains to show that U_2 and T_2 are of a smaller order than U_1 in probability. For U_2, we have

$$U_2 = \frac{1}{2n}\sum_{i=1}^{n}\left\{\alpha_m^2\boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{u}_{2i} + 2\alpha_m\boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_{0i}\right\}.$$

Focusing on the second term in U_2, let $\boldsymbol{h}_i = \boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}$. Then by the Cauchy-Schwarz inequality, $\boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_{0i} \le \|\boldsymbol{h}_i\|\times\|\boldsymbol{b}_{0i}\| \le C\|\boldsymbol{b}_{0i}\|$ for some sufficiently large constant C. Since $\boldsymbol{b}_{0il} = 0$ for $l = p_{0r}+1, \dots, p_r$, we need only consider the subvector $\check{\boldsymbol{b}}_{0i}$ consisting of the first p_{0r} elements of $\boldsymbol{b}_{0i}$. Moreover, $\check{\boldsymbol{b}}_{0i}$ is drawn from a multivariate normal distribution with zero mean vector and positive definite covariance matrix $\boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{01}^T$. Write $\boldsymbol{\Gamma}_{01}\boldsymbol{\Gamma}_{01}^T = \boldsymbol{Q}_{01}\boldsymbol{\Lambda}_{01}\boldsymbol{Q}_{01}^T$ where $\boldsymbol{Q}_{01}$ is an orthogonal matrix of eigenvectors, and $\boldsymbol{\Lambda}_{01}$ is a diagonal matrix consisting of p_{0r} positive eigenvalues. Using this result, we obtain $\|\boldsymbol{b}_{0i}\| = \|\check{\boldsymbol{b}}_{0i}\| = \|\boldsymbol{Q}_{01}^T\check{\boldsymbol{b}}_{0i}\|$. Then by the version of Chebychev's inequality presented in Ferentinos (1982), we have

$$P\left(\|\boldsymbol{Q}_{01}^T\check{\boldsymbol{b}}_{0i}\| \ge \log(m_1)\left\{\mathrm{tr}(\boldsymbol{\Lambda}_{01}^T\boldsymbol{\Lambda}_{01})\right\}^{1/2}\right) \le \frac{1}{\log(m_1)^2} \to 0.$$

It follows that as $m_1 \to \infty$, $\alpha_m\boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_{0i}$ is bounded above by

$\alpha_m K \log(m_1)\{\mathrm{tr}(\boldsymbol{\Lambda}_{01}^T\boldsymbol{\Lambda}_{01})\}$ in probability. Thus we conclude that $\alpha_m \boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_{0i} = O_p\{\alpha_m \log(m_1)\} = o_p(1)$. Turning to the first term in U_2, let $\tau_{\max}\{(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\}$ denote the maximum eigenvalue of $(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}$, and note that it satisfies $\tau_{\max}\{(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\} < \infty$ for all $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. It follows that $\alpha_m^2 \boldsymbol{u}_{2i}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{u}_{2i} \leq \alpha_m^2\|\boldsymbol{u}_{2i}\|^2\tau_{\max}\{(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\} \leq \alpha_m^2\|\boldsymbol{u}_{2i}\|^2 \max_r \left(\sum_s \left|(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}_{rs}\right|\right) = O_p(\alpha_m^2) = o_p(1)$ where $(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}_{rs}$ denotes element (r,s) of $(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}$. Combining the results above, we have $U_2 = o_p(1)$ and conclude that U_2 is asymptotically dominated by U_1.

Turning to term T_2, note that for any real number c, $(b_{0il} + \alpha_m c)^2 \geq b_{0il}^2 - 2\alpha_m|c| \geq 0$ for all m_i sufficiently large. Therefore,

$$\begin{aligned}\frac{1}{n}\left(\|\boldsymbol{b}_{0\bullet l} + \alpha_m \boldsymbol{u}_{2l}\| - \|\boldsymbol{b}_{0\bullet l}\|\right) &\geq \frac{1}{n}\left\{\sum_{i=1}^n \left(b_{0il}^2 - 2\alpha_m|u_{2il}|\right)\right\}^{1/2} - \frac{1}{n}\|\boldsymbol{b}_{0\bullet l}\| \\ &= \frac{1}{n}\|\boldsymbol{b}_{0\bullet l}\|\left\{(1-\delta)^{1/2} - 1\right\},\end{aligned}$$

where $\delta = 2\alpha_m \sum_{i=1}^n |u_{2il}| \left(\sum_{i=1}^n b_{0il}^2\right)^{-1} \xrightarrow{p} 0$. We can then apply a Taylor expansion of $(1-\delta)^{1/2}$ about $\delta = 0$ to obtain $(1-\delta)^{1/2} - 1 = -2^{-1}\delta + O_p(\delta^2)$. Following some simplification, we have that for all m_i large enough,

$$-T_2 \leq \frac{\lambda}{n}\left\{\sum_{k=2}^{p_{0f}} v_k \alpha_m |u_{1k}| \mathrm{sgn}(u_{1k}\beta_{0k}) + \alpha_m \sum_{l=1}^{p_{0r}}\left(\frac{w_l \sum_{i=1}^n |u_{2il}|}{\|\boldsymbol{b}_{0\bullet l}\|}\right)\right\}.$$

For all $k = 2, \ldots, p_{0f}$ and $l = 1, \ldots, p_{0r}$, it holds that $v_k = O_p(1)$ and $w_l = O_p(1)$ by the design of the adaptive lasso weights. Therefore, we obtain

$$\frac{\lambda\alpha_m}{n}\sum_{k=2}^{p_{0f}} v_k|u_{1k}|\mathrm{sgn}(u_{1k}\beta_{0k}) \leq O_p\left(\frac{\lambda\alpha_m}{n}\right)\|\boldsymbol{u}_1\| \leq O_p\left(\frac{\lambda\alpha_m}{n}\right)$$

and

$$\frac{\lambda\alpha_m}{n}\sum_{l=1}^{p_{0r}}\left(\frac{w_l\sum_{i=1}^{n}|u_{2il}|}{\|\boldsymbol{b}_{0\bullet l}\|}\right)\le O_p\left(\frac{\lambda\alpha_m}{n^{1/2}}\right).$$

By condition (C5a) then, $-T_2 \le O_p\left(\lambda\alpha_m n^{-1/2}\right) = o_p(Nm_1^{-1})$. Hence for all m_i and n large enough, T_2 is also dominated by U_1 in probability. It follows that for a large enough constant K_2, we have $P\left(\sup_{\|\boldsymbol{u}\|=n^{1/2}K_2}\Delta(\boldsymbol{u},\boldsymbol{\Gamma})<0\right)\to 1$. Since $\ell_p(\boldsymbol{\Psi},\boldsymbol{b})$ is strictly concave, replacing $\boldsymbol{\Gamma}$ with the estimator based on equation (2) in the main text, it follows that the regularized PQL estimator exists and falls within a $m_1^{-1/2}$-neighborhood of $(\boldsymbol{\beta}_0,\boldsymbol{b}_0)$, and the result follows.

It is interesting to note that the $m_1^{1/2}$-consistency holds irrespective of the value of $\boldsymbol{\Gamma}$. This result ties in with previous work of Jiang et al. (2001), who showed that in certain large sample situations, the consistency of the fixed and random effect coefficients is unaffected by the variance components at which they are estimated.

Finally, note the consistency result for $\hat{\boldsymbol{b}}_{\lambda i}$ implies that

$$\left(\sum_{i=1}^{n}\|\hat{\boldsymbol{b}}_{\lambda i}-\boldsymbol{b}_{0i}\|^2\right)^{1/2}=\|\hat{\boldsymbol{b}}_{\lambda}-\boldsymbol{b}_0\|=O_p\left\{\left(\frac{n}{m_1}\right)^{1/2}\right\},$$

a result that we shall make use of in the proof of selection consistency later on. □

2 Proof of selection consistency in Theorem 1

For a given $\boldsymbol{D}$, the regularized PQL estimator is given by

$$(\hat{\boldsymbol{\beta}}_\lambda,\hat{\boldsymbol{b}}_\lambda)=\arg\max_{\boldsymbol{\beta},\boldsymbol{b}}\ell_p(\boldsymbol{\Psi},\boldsymbol{b})=\arg\max_{\boldsymbol{\beta},\boldsymbol{b}}\ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})-\lambda\sum_{k=1}^{p_f}v_k|\beta_k|-\lambda\sum_{l=1}^{p_r}w_l\|\boldsymbol{b}_{\bullet l}\|.$$

Let $\hat{\boldsymbol{\Psi}}_\lambda = (\hat{\boldsymbol{\beta}}_\lambda^T, \hat{\boldsymbol{\Gamma}}_\lambda^T)^T$ where $\hat{\boldsymbol{D}}_\lambda = \hat{\boldsymbol{\Gamma}}_\lambda \hat{\boldsymbol{\Gamma}}_\lambda^T$ is estimated using the iterative formula in (2) in the main text.

We split the proof of selection consistency into separate arguments for the fixed and random effect coefficients. Given the regularized PQL estimators are $m_1^{1/2}$-consistent from Theorem 1, it suffices to show that for all truly zero coefficients, the sign of the score equation depends on the sign of the estimated coefficient with probability tending to one.

<u>Fixed effects:</u> We have $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0 = O_p\left(m_1^{-1/2}\right)$ from Theorem 1. Suppose now that for some $k = 2, \ldots, p_f$ with $\beta_{0k} = 0$, we have $\hat{\beta}_{\lambda k} \neq 0$. By a Taylor expansion of the score equation about $(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)$,

$$\begin{aligned}\left.\frac{\partial \ell_p(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k}\right|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda} &= \left.\frac{\partial \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k}\right|_{\boldsymbol{\Psi}_0, \boldsymbol{b}_0} + \left.\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k \partial \boldsymbol{\Psi}}\right|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \left(\hat{\boldsymbol{\Psi}}_\lambda - \boldsymbol{\Psi}_0\right) \\ &\quad + \left.\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k \partial \boldsymbol{b}}\right|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \left(\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\right) - \lambda v_k \text{sgn}(\hat{\beta}_{\lambda k}) \\ &\triangleq T_1 + T_2 + T_3 + T_4,\end{aligned}$$

where $(\bar{\boldsymbol{\Psi}}^T, \bar{\boldsymbol{b}}^T)^T$ lies on the line segment joining $(\hat{\boldsymbol{\Psi}}_\lambda^T, \hat{\boldsymbol{b}}_\lambda^T)^T$ and $(\boldsymbol{\Psi}_0^T, \boldsymbol{b}_0^T)^T$. Let $\boldsymbol{X}_{ik}$ denote the k^{th} column of $\boldsymbol{X}_i$. Then from the proof of estimation consistency, we have

$$T_1 = \left.\frac{\partial \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k}\right|_{\boldsymbol{\Psi}_0, \boldsymbol{b}_0} = \sum_{i=1}^n \boldsymbol{X}_{ik}^T \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0, \boldsymbol{b}_{0i})\} = O_p(N^{1/2}).$$

Next, by conditions (C1)-(C2), we obtain

$$-\frac{1}{n} \left.\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k \partial \beta_l}\right|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_{ik}^T \boldsymbol{W}_i(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{b}}) \boldsymbol{X}_{il} = O_p\left(\frac{N}{n}\right); \quad l = 1, \ldots, p_f.$$

Consider now the iterative update estimator of the random effect covariance matrix in equation (2) of the main text, $\hat{\boldsymbol{D}}_\lambda^{(t)} = n^{-1}\sum_{i=1}^n \left\{ \left(\boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_{\lambda i} \boldsymbol{Z}_i + (\hat{\boldsymbol{D}}_\lambda^{(t-1)})^- \right)^{-1} + \hat{\boldsymbol{b}}_{\lambda i}\hat{\boldsymbol{b}}_{\lambda i}^T \right\}$. As every $m_i \to \infty$, $\boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_{\lambda i} \boldsymbol{Z}_i$ asymptotically dominates $(\hat{\boldsymbol{D}}_\lambda^{(t-1)})^-$, and so by conditions (C1)-(C2) we have $\left(\boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_{\lambda i} \boldsymbol{Z}_i + (\hat{\boldsymbol{D}}_\lambda^{(t-1)})^- \right)^{-1} = O_p(m_1^{-1})$. Finally, since $\hat{\boldsymbol{b}}_{\lambda i}$ is $m_1^{1/2}$-consistent from the proof of Theorem 1 for all $i = 1, \ldots, n$, then $\hat{\boldsymbol{b}}_{\lambda i}\hat{\boldsymbol{b}}_{\lambda i}^T$ asymptotically dominates $\left(\boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_{\lambda i} \boldsymbol{Z}_i + (\hat{\boldsymbol{D}}_\lambda^{(t-1)})^- \right)^{-1}$ in probability, and we conclude that $\hat{\boldsymbol{D}}_\lambda^{(t)}$ is also a $m_1^{1/2}$-consistent estimator of the true covariance matrix. It follows that $\text{vec}(\hat{\boldsymbol{\Gamma}}_\lambda) - \text{vec}(\boldsymbol{\Gamma}_0) = O_p(m_1^{-1/2})$ holds, where $\hat{\boldsymbol{D}}_\lambda = \hat{\boldsymbol{\Gamma}}_\lambda \hat{\boldsymbol{\Gamma}}_\lambda^T$.

Since $\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \text{vec}(\boldsymbol{\Gamma})^T)^T$, and $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\| = O_p(m_1^{-1/2})$ by Theorem 1, then $\|\hat{\boldsymbol{\Psi}}_\lambda - \boldsymbol{\Psi}_0\| = O_p(m_1^{-1/2})$. Hence by the Cauchy-Schwarz inequality, and noting that $\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b}) / \partial\beta_k \partial \text{vec}(\boldsymbol{\Gamma})_r = 0$ for all $r = 1, \ldots, \dim\{\text{vec}(\boldsymbol{\Gamma})\}$, we have

$$\begin{aligned} |T_2| &\leq n \left\{ \sum_{l=1}^{p_f} \left(\frac{1}{n} \frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial\beta_k \partial\beta_l} \bigg|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \right)^2 + \left(\sum_{r=1}^{\dim\{\text{vec}(\boldsymbol{\Gamma})\}} \frac{1}{n} \frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial\beta_k \partial \text{vec}(\boldsymbol{\Gamma})_r} \bigg|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \right)^2 \right\}^{1/2} \|\hat{\boldsymbol{\Psi}}_\lambda - \boldsymbol{\Psi}_0\| \\ &= O_p\left(\frac{N}{m_1^{1/2}} \right), \end{aligned}$$

where $\text{vec}(\boldsymbol{\Gamma})_r$ is the r^{th} element of $\text{vec}(\boldsymbol{\Gamma})$. For T_3, we have for all $i = 1, \ldots, n$ and $l = 1, \ldots, p_r$,

$$-\frac{1}{m_i} \frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial\beta_k \partial b_{il}} \bigg|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} = \frac{1}{m_i} \boldsymbol{X}_{ik}^T \boldsymbol{W}_i(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{b}}) \boldsymbol{Z}_{il} = O_p(1),$$

where $\boldsymbol{Z}_{il}$ denotes the l^{th} column of $\boldsymbol{Z}_i$. Using the previously obtained result that $\|\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\| = O_p(n^{1/2} m_1^{-1/2})$, we have

$$|T_3| \leq \left\{ \sum_{i=1}^n \sum_{l=1}^{p_r} \left(m_i \times \frac{1}{m_i} \frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial\beta_k \partial b_{il}} \bigg|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \right)^2 \right\}^{1/2} \|\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\|$$

$$
\begin{aligned}
&= O_p\left\{\left(\sum_{i=1}^{n} m_i^2\right)^{1/2}\right\} \times O_p\left\{\left(\frac{n}{m_1}\right)^{1/2}\right\} \\
&= O_p\left(\frac{N}{m_1^{1/2}}\right),
\end{aligned}
$$

where we have used the fact that $N = \sum_{i=1}^{n} m_i = O_p(nm_n)$. Combining the results obtained for terms T_1 to T_3 then, we obtain

$$
\left.\frac{\partial \ell_p(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial \beta_k}\right|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda} = O_p\left(\frac{N}{m_1^{1/2}}\right) - \lambda v_k \mathrm{sgn}(\hat{\beta}_{\lambda k}) = \lambda v_k \left\{O_p\left(\frac{N}{\lambda v_k m_1^{1/2}}\right) - \mathrm{sgn}(\hat{\beta}_{\lambda k})\right\},
$$

where $\lambda v_k > 0$. Since $\beta_{0k} = 0$, then $v_k = |\tilde{\beta}_k|^{-\kappa} = O_p(n^{\kappa/2})$. Therefore, by condition (C5b), it follows that with probability tending to one, the sign of the score equation is completely determined by the sign of $\hat{\beta}_{\lambda k}$, and we conclude that $P(\hat{\beta}_{\lambda k} = 0) \to 1$ for all k with $\beta_{0k} = 0$.

Random effects: Suppose for some $l = 1, \ldots, p_r$ with $\boldsymbol{\Gamma}_{0,lq} = 0$ for all $q = 1, \ldots, p_r$, we have $\hat{b}_{\lambda il} \neq 0$. Note that $\boldsymbol{\Gamma}_{0,lq} = 0$ implies $\|\boldsymbol{b}_{0\bullet l}\| = 0$. Similar to the proof for the fixed effects, consider a Taylor expansion of the regularized PQL score equation about $(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)$,

$$
\begin{aligned}
\left.\frac{\partial \ell_p(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial b_{il}}\right|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda} &= \left.\frac{\partial \ell_{\mathrm{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial b_{il}}\right|_{\boldsymbol{\Psi}_0, \boldsymbol{b}_0} + \left.\frac{\partial^2 \ell_{\mathrm{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial b_{il} \partial \boldsymbol{\Psi}}\right|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \left(\hat{\boldsymbol{\Psi}}_\lambda - \boldsymbol{\Psi}_0\right) \\
&\quad + \left.\frac{\partial^2 \ell_{\mathrm{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial b_{il} \partial \boldsymbol{b}}\right|_{\bar{\boldsymbol{\Psi}}, \bar{\boldsymbol{b}}} \left(\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\right) - \lambda w_l \frac{\hat{b}_{\lambda il}}{\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\|} \\
&\triangleq T_1 + T_2 + T_3 + T_4,
\end{aligned}
$$

where $(\bar{\boldsymbol{\Psi}}^T, \bar{\boldsymbol{b}}^T)^T$ lies on the line segment joining $(\hat{\boldsymbol{\Psi}}_\lambda^T, \hat{\boldsymbol{b}}_\lambda^T)^T$ and $(\boldsymbol{\Psi}_0^T, \boldsymbol{b}_0^T)^T$. From the proof of estimation consistency, we have $T_1 = \boldsymbol{Z}_{il}^T\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0, \boldsymbol{b}_{0i})\} - (\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)^{-}\boldsymbol{b}_{0i} =$

$O_p(m_i^{1/2})$. For T_2, we have by conditions (C1)-(C2),

$$-\frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial\beta_k}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}} = \frac{1}{m_i}\boldsymbol{Z}_{il}^T\boldsymbol{W}_i(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}})\boldsymbol{X}_{ik} = O_p(1); \quad k=1,\ldots,p_f$$
$$-\frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial \text{vec}(\boldsymbol{\Gamma})}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}} = \frac{1}{m_i}\frac{\partial[(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_i]_l}{\partial\text{vec}(\boldsymbol{\Gamma})}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}} = O_p(m_i^{-1}),$$

where $[(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_i]_l$ refers to the l^{th} element of $(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-}\boldsymbol{b}_i$. Hence by the Cauchy-Schwarz inequality,

$$\begin{aligned}|T_2| &\leq \left\{\sum_{k=1}^{p_f}\left(m_i \times \frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial\beta_k}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}}\right)^2 + \left(\sum_{r=1}^{\dim\{\text{vec}(\boldsymbol{\Gamma})\}} m_i \times \frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial\text{vec}(\boldsymbol{\Gamma})_r}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}}\right)^2\right\}^{1/2} \\ &\quad \times \|\hat{\boldsymbol{\Psi}}_\lambda - \boldsymbol{\Psi}_0\| \\ &= O_p\left(\frac{m_n}{m_1^{1/2}}\right).\end{aligned}$$

Recall that m_n is the cluster size with the largest rate of growth. For T_3, we have

$$-\frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial b_{iq}}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}} = \frac{1}{m_i}\boldsymbol{Z}_{il}^T\boldsymbol{W}_i(\bar{\boldsymbol{\beta}},\bar{\boldsymbol{b}})\boldsymbol{Z}_{iq} + (\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{\Gamma}}^T)^{-}_{lq} = O_p(1); \quad q=1,\ldots,p_r$$
$$-\frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial b_{jq}} = 0; \quad j \neq i;\ q=1,\ldots,p_r.$$

where $(\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{\Gamma}}^T)^{-}_{lq}$ denotes element (l,q) of $(\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{\Gamma}}^T)^{-}$. Along with $\|\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\| = O_p(n^{1/2}m_1^{-1/2})$, the above results imply

$$|T_3| \leq \left\{\sum_{i=1}^{n}\sum_{q=1}^{p_r}\left(m_i \times \frac{1}{m_i}\frac{\partial^2 \ell_{\text{PQL}}(\boldsymbol{\Psi},\boldsymbol{b})}{\partial b_{il}\partial b_{iq}}\bigg|_{\bar{\boldsymbol{\Psi}},\bar{\boldsymbol{b}}}\right)^2\right\}^{1/2} \|\hat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_0\|$$

$$
\begin{aligned}
&= O_p\left\{\left(\sum_{i=1}^{n} m_i^2\right)^{1/2}\right\} \times O_p\left\{\left(\frac{n}{m_1}\right)^{1/2}\right\} \\
&= O_p\left(\frac{N}{m_1^{1/2}}\right).
\end{aligned}
$$

Combining the results for T_1 to T_3, we obtain

$$
\begin{aligned}
\left.\frac{\partial \ell_p(\boldsymbol{\Psi}, \boldsymbol{b})}{\partial b_{il}}\right|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda} &= O_p\left(\frac{N}{m_1^{1/2}}\right) - \lambda w_l \mathrm{sgn}(\hat{b}_{\lambda il}) \frac{|\hat{b}_{\lambda il}|}{\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\|} \\
&= \lambda w_l \left\{O_p\left(\frac{N}{\lambda w_l m_1^{1/2}}\right) - \mathrm{sgn}(\hat{b}_{\lambda il}) \frac{|\hat{b}_{\lambda il}|}{\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\|}\right\},
\end{aligned}
$$

where $\lambda w_l > 0$. Note that $|\hat{b}_{\lambda il}|/\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\| \in (0, 1)$ for every n and λ. By the construction of the adaptive lasso weights, $\boldsymbol{\Gamma}_{0,lq} = 0$ implies that $w_l = \tilde{\boldsymbol{D}}_{ll}^{-\kappa} = O_p(n^{\kappa/2})$. Therefore, by condition (C5b), with probability tending to one, the sign of the score equation is completely determined by the sign of $\hat{b}_{\lambda il}$. We conclude that $P(\hat{b}_{\lambda il} = 0) \to 1$ for all $i = 1, \ldots, n$ and l such that $\|\boldsymbol{b}_{0 \bullet l}\| = 0$, and hence $P(\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\| = 0) \to 1$. □

3 Proof of Lemma 1

We note that for any submodel α chosen by a tuning parameter λ, a similar argument to Theorem 1 can be used to show that the unregularized PQL estimates satisfy

$$
\begin{aligned}
&\|\tilde{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_\alpha^*\| = O_p(m_1^{-1/2}) \\
&\|\tilde{\boldsymbol{b}}_{\alpha i} - \boldsymbol{b}_{\alpha i}^*\| = O_p(m_1^{-1/2}) \quad \text{for all} \quad i = 1, \ldots, n,
\end{aligned} \tag{3}
$$

where $(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)$ are the pseudo-true parameters for submodel α, defined as $\boldsymbol{\Psi}_\alpha^* = \{\boldsymbol{\beta}_\alpha^{*T}, \mathrm{vech}(\boldsymbol{D})_\alpha^{*T}\}^T = \arg\min_{\boldsymbol{\Psi}_\alpha} E\{-\ell(\boldsymbol{\Psi}_\alpha)\}$, where $\ell(\boldsymbol{\Psi}_\alpha)$ is the marginal-likelihood for model α, and $\boldsymbol{b}_\alpha^*$ is a realization from the a multivariate normal distribution with mean vector zero and covariance $\boldsymbol{D}_\alpha^*$ (see White, 1982, for more detail on the concept of pseudo-true parameters).

The proof of (3) is straightforward: consider any submodel α that defines a subset of the fixed and random covariates. Suppose now that we treat this as the new full model, and perform regularized PQL estimation using another tuning parameter, say, λ_2. Now, the *unregularized* PQL estimator of the model α corresponds to the case $\lambda_2 = 0$. Since $\lambda_2 = 0$ satisfies condition (C5a), the first part of Theorem 1 can be applied to this submodel to obtain $m_1^{1/2}$-consistency for $\tilde{\boldsymbol{\beta}}_\alpha$ and the $\tilde{\boldsymbol{b}}_{\alpha i}$'s to the pseudo-true parameters.

It is important to recognize that the pseudo-true parameters are not necessarily equal to the true parameters. If the submodel α underfits, i.e. fails to contain at least one of the truly non-zero fixed and/or random effects, then the $(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)$ will not correspond to (the elements of) the true parameter point. This is represented by condition (C6) in the main text. On the other hand, if the submodel α overfits, i.e. contains all the truly non-zero fixed and/or random effects as well as one or more truly zero effect, then $\boldsymbol{\beta}_\alpha^*$ will be a subvector of $\boldsymbol{\beta}_0$ and $\boldsymbol{b}_{\alpha i}^*$ will be a subvector of $\boldsymbol{b}_{0i}$.

Note that for the tuning parameter λ_0 satisfying conditions (C5), such that $\alpha = \alpha_0$, we have $\boldsymbol{\beta}_{\alpha_0}^* = \boldsymbol{\beta}_{01}$ and $\boldsymbol{b}_{\alpha_0 i}^* = \boldsymbol{b}_{01i}$ for all $i = 1, \ldots, n$, where $\boldsymbol{b}_{01i}$ contains the first p_{0r} elements of $\boldsymbol{b}_{0i}$.

We treat separately the two cases of underfitted and overfitted models.

<u>Underfitted models:</u> Here, we have $\alpha \not\supset \alpha_0$. Let $D = N^{-1}\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0})\}$, where $\ell_1(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \log f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)$. Then from the definition of $\text{IC}_{\text{proxy}}(\alpha)$ in Section 4.1 of the main text,

$$\text{IC}_{\text{proxy}}(\alpha) - \text{IC}_{\text{proxy}}(\alpha_0) \geq -2D - \frac{\log(n)}{N}\dim(\tilde{\boldsymbol{\beta}}_{\alpha_0}) - \frac{2}{N}\dim(\tilde{\boldsymbol{b}}_{\alpha_0}), \tag{4}$$

after dropping the non-negative terms. Since the second and third terms on the right hand side of (4) are of order $o_p(1)$, we focus only on the first term. By definition of the unregularized PQL estimates, $\ell_{\text{PQL}}(\tilde{\boldsymbol{\Psi}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) \geq \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)$. Therefore, we can write

$$
\begin{aligned}
-2D &= -\frac{2}{N}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_{\text{PQL}}(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) - \frac{1}{2}\sum_{i=1}^{n} \tilde{\boldsymbol{b}}_{\alpha_0 i}^T \tilde{\boldsymbol{D}}_{\alpha_0}^{-} \tilde{\boldsymbol{b}}_{\alpha_0 i}\right\} \\
&\geq -\frac{2}{N}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0) - \frac{1}{2}\sum_{i=1}^{n} \tilde{\boldsymbol{b}}_{\alpha_0 i}^T \tilde{\boldsymbol{D}}_{\alpha_0}^{-} \tilde{\boldsymbol{b}}_{\alpha_0 i}\right\} \\
&= -\frac{2}{N}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_\alpha^*, \boldsymbol{b}_\alpha^*)\right\} \\
&\quad - \frac{2}{N}\left\{\ell_{\text{PQL}}(\boldsymbol{\Psi}_\alpha^*, \boldsymbol{b}_\alpha^*) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0) - \frac{1}{2}\sum_{i=1}^{n} \tilde{\boldsymbol{b}}_{\alpha_0 i}^T \tilde{\boldsymbol{D}}_{\alpha_0}^{-} \tilde{\boldsymbol{b}}_{\alpha_0 i}\right\} \\
&= -\frac{2}{N}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)\right\} \\
&\quad - \frac{2}{N}\left\{\ell_{\text{PQL}}(\boldsymbol{\Psi}_\alpha^*, \boldsymbol{b}_\alpha^*) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)\right\} + O_p\left(\frac{n}{N}\right) \\
&\triangleq T_1 + T_2 + o_p(1)
\end{aligned}
$$

where $\boldsymbol{b}_0 = (\boldsymbol{b}_{01}^T, \ldots, \boldsymbol{b}_{0n}^T)^T$ and the order term $O_p(nN^{-1})$ in the second last line of the above comes from noting that $\sum_{i=1}^{n} \tilde{\boldsymbol{b}}_{\alpha_0 i}^T \tilde{\boldsymbol{D}}_{\alpha_0}^{-} \tilde{\boldsymbol{b}}_{\alpha_0 i} = O_p(n)$. From Vonesh (1996), we have $n^{-1}\ell(\boldsymbol{\Psi}_0) = n^{-1}\ell_{\text{LA}}(\boldsymbol{\Psi}_0) + O_p(m_1^{-1})$ where $\ell(\boldsymbol{\Psi})$ and $\ell_{\text{LA}}(\boldsymbol{\Psi})$ are the marginal log-likelihood and the Laplace approximated log-likelihood respectively. Furthermore from Section 7.3 of Demidenko (2013) for instance, we can show that $n^{-1}\ell_{\text{LA}}(\boldsymbol{\Psi}_0) = n^{-1}\ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0) + o_p(1)$ as $m_1/n \to 0$. Combining these two results, we obtain $n^{-1}\left|\ell(\boldsymbol{\Psi}_0) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)\right| \xrightarrow{p} 0$. Since $n^{-1}\ell(\boldsymbol{\Psi}_0) \xrightarrow{p} \text{E}\{\ell_o(\boldsymbol{\Psi}_0)\}$ by the weak law of large numbers, then we have $n^{-1}\ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0) \xrightarrow{p} \text{E}\{\ell_o(\boldsymbol{\Psi}_0)\}$. In a similar manner, for any model α we can show that $n^{-1}\ell_{\text{PQL}}(\boldsymbol{\Psi}_\alpha^*, \boldsymbol{b}_\alpha^*) \xrightarrow{p} \text{E}\{\ell_o(\boldsymbol{\Psi}_\alpha^*)\}$. We therefore have

$$
-\frac{2}{n}\left\{\ell_{\text{PQL}}(\boldsymbol{\Psi}_\alpha^*, \boldsymbol{b}_\alpha^*) - \ell_{\text{PQL}}(\boldsymbol{\Psi}_0, \boldsymbol{b}_0)\right\} \xrightarrow{p} 2\text{E}\left\{\ell_o(\boldsymbol{\Psi}_0) - \ell_o(\boldsymbol{\Psi}_\alpha^*)\right\},
$$

and by condition (C6), the limit on the right hand side exceeds $c_2 > 0$ for all underfitted models. It follows immediately that term T_2 is asymptotically greater than zero, and it remains to prove that T_1 is asymptotically negligible.

Let $\boldsymbol{v} = (\tilde{\boldsymbol{\beta}}_\alpha^T - \boldsymbol{\beta}_\alpha^{*T}, \tilde{\boldsymbol{b}}_\alpha^T - \boldsymbol{b}_\alpha^{*T})^T$. Applying a Taylor expansion to $\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha)$, we obtain

$$
\begin{aligned}
-\frac{2}{N}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)\right\} &= -\frac{2}{N}\boldsymbol{v}^T \nabla \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*) \\
&+ \boldsymbol{v}^T\left\{-\frac{1}{N}\nabla^2 \ell_1(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)\right\}\boldsymbol{v}, \qquad (5)
\end{aligned}
$$

where $(\bar{\boldsymbol{\beta}}_\alpha^T, \bar{\boldsymbol{b}}_\alpha^T)^T$ lies on the line segment joining $(\tilde{\boldsymbol{\beta}}_\alpha^T, \tilde{\boldsymbol{b}}_\alpha^T)^T$ and $(\boldsymbol{\beta}_\alpha^{*T}, \boldsymbol{b}_\alpha^{*T})^T$. For the first term on the right hand side equation (5), we have

$$
\nabla \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*) = \left[\sum_{i=1}^n \boldsymbol{X}_i^T\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_{\alpha i}^*)\}, \boldsymbol{Z}_1^T\{\boldsymbol{y}_1 - \boldsymbol{\mu}_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_{\alpha 1}^*)\}, \ldots, \boldsymbol{Z}_n^T\{\boldsymbol{y}_n - \boldsymbol{\mu}_n(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_{\alpha n}^*)\}\right].
$$

Similar to equation (1) and the proof of estimation consistency in Theorem 1, we can show $\sum_{i=1}^n \boldsymbol{X}_i^T\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_{\alpha i}^*)\} = O_p\{N^{1/2}\}$ and $\boldsymbol{Z}_i^T\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_{\alpha i}^*)\} = O_p(m_i^{1/2})$ for all $i = 1, \ldots, n$. Furthermore by the $m_1^{1/2}$-consistency of the unregularized PQL estimates for the pseudo-true parameters, as discussed at the beginning of this proof, we have $\|\boldsymbol{v}\| = O_p(n^{1/2} m_1^{-1/2})$. By the Cauchy-Schwarz inequality, we have

$$
\left\|\frac{1}{N}\boldsymbol{v}^T \nabla \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)\right\| \leq \frac{1}{N}\|\boldsymbol{v}\| \, \|\nabla \ell_1(\boldsymbol{\beta}_\alpha^*, \boldsymbol{b}_\alpha^*)\| = O_p(n^{1/2} N^{-1/2}) = o_p(1).
$$

Next, equation (2) describes the structure of $\nabla^2 \ell_1(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. By applying conditions (C1)-(C2) and (C4), and the $m_1^{1/2}$-consistency of $\tilde{\boldsymbol{\beta}}_\alpha$ and $\tilde{\boldsymbol{b}}_{\alpha i}$, we obtain the following results: $N^{-1}\sum_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{W}_i(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_{\alpha i})\boldsymbol{X}_i = O_p(1)$ and for all $i = 1, \ldots, n$,

$$
\frac{1}{N}\boldsymbol{Z}_i^T \boldsymbol{W}_i(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_{\alpha i})\boldsymbol{X}_i = O_p\left(\frac{m_i}{N}\right) \quad \text{and} \quad \frac{1}{N}\boldsymbol{Z}_i^T \boldsymbol{W}_i(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_{\alpha i})\boldsymbol{Z}_i = O_p\left(\frac{m_i}{N}\right).
$$

Using these results and doing straightforward matrix algebra with the orders of the various quantities, we can show that $\boldsymbol{v}^T \left\{-N^{-1}\nabla^2\ell_1(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)\right\} \boldsymbol{v} = O_p(m_1^{-1}) = o_p(1)$. Combining the above results, we have $T_1 = -2N^{-1}\left\{\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\boldsymbol{\beta}^*_\alpha, \boldsymbol{b}^*_\alpha)\right\} = o_p(1)$, and therefore $-2N^{-1}D > c_2 + o_p(1)$. As this holds for all underfitted models, we conclude that the left hand side of (4) is guaranteed to be positive with probability tending to one.

Overfitted models: Here, $\alpha \supset \alpha_0$, i.e. α is a proper superset of α_0. Let $d_f = \dim(\tilde{\boldsymbol{\beta}}_\alpha) - \dim(\tilde{\boldsymbol{\beta}}_{\alpha_0}) \geq 0$ and $d_r = \dim(\tilde{\boldsymbol{b}}_{\alpha i}) - \dim(\tilde{\boldsymbol{b}}_{\alpha_0 i}) \geq 0$ denote the number of overfitted fixed and random effects respectively. By the definition of overfitting, at least one of d_f and d_r must be positive. From the definition of $\mathrm{IC}_{\text{proxy}}(\alpha)$, we have

$$N\left\{\mathrm{IC}_{\text{proxy}}(\alpha) - \mathrm{IC}_{\text{proxy}}(\alpha_0)\right\} = -2D + \log(n)d_f + 2nd_r, \tag{6}$$

where $D = \ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0})$. Adding and subtracting $\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})$, where we recall that $\boldsymbol{\beta}_{01}$ contains all the non-zero elements for $\boldsymbol{\beta}_0$, the vector $\boldsymbol{b}_{01i}$ contains all the non-zero elements of $\boldsymbol{b}_{0i}$ for $i = 1, \ldots, n$, and $\boldsymbol{b}_{01} = (\boldsymbol{b}_{011}^T, \ldots, \boldsymbol{b}_{01n}^T)^T$, we have

$$\begin{aligned}|D| &\leq \left|\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) - \ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})\right| + \left|\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) - \ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})\right| \\ &= |E_1| + |E_2|.\end{aligned}$$

For E_1, let $\boldsymbol{v}_\beta = \tilde{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_{01}$, where the appropriate augmentation has been performed to ensure $\boldsymbol{\beta}_{01}$ is of the same dimension as the overfitted coefficients $\tilde{\boldsymbol{\beta}}_\alpha$. Similarly, define $\boldsymbol{v}_b$ as the difference between $\tilde{\boldsymbol{b}}_\alpha$ and $\boldsymbol{b}_{01}$. Letting $\boldsymbol{H}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha) = -N^{-1}\nabla^2\ell_1(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)$, then applying a Taylor expansion to $\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha)$ we obtain

$$E_1 = (\boldsymbol{v}_\beta^T, \boldsymbol{v}_b^T)^T\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01}) - \frac{N}{2}(\boldsymbol{v}_\beta^T, \boldsymbol{v}_b^T)^T\boldsymbol{H}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)(\boldsymbol{v}_\beta^T, \boldsymbol{v}_b^T),$$

where $(\bar{\boldsymbol{\beta}}_\alpha^T, \bar{\boldsymbol{b}}_\alpha^T)^T$ lies on the line segment joining $(\tilde{\boldsymbol{\beta}}_\alpha^T, \tilde{\boldsymbol{b}}_\alpha^T)^T$ and $(\boldsymbol{\beta}_{01}^T, \boldsymbol{b}_{01}^T)^T$, with the appropriate augmentation performed implicitly. Now, note that for the overfitted model α, we can apply equation (3) to show that $\|(\tilde{\boldsymbol{\beta}}_\alpha^T, \tilde{\boldsymbol{b}}_\alpha^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{b}_0^T)^T\|_\infty \leq \varepsilon$ for n and m_1 large enough. Therefore, by condition (C3), $\boldsymbol{H}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)$ is invertible and we can utilize the result

$$
\begin{aligned}
-\frac{N}{2}\left\{(\boldsymbol{v}_\beta^T, \boldsymbol{v}_b^T) - N^{-1}\boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})\right\}^T \boldsymbol{H}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha) \\
\times\left\{(\boldsymbol{v}_\beta^T, \boldsymbol{v}_b^T) - N^{-1}\boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})\right\} \leq 0,
\end{aligned}
$$

to obtain

$$
E_1 \leq \frac{1}{N}\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})^T\boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01}).
$$

By condition (C3), $\boldsymbol{H}(\bar{\boldsymbol{\beta}}_\alpha, \bar{\boldsymbol{b}}_\alpha)$ has a minimum eigenvalue $\tau_{\min}$ that is bounded away from zero. Therefore we obtain $E_1 \leq (N\tau_{\min})^{-1}\|\nabla\ell_1(\boldsymbol{\beta}_{01}, \boldsymbol{b}_{01})\|^2$, and from equation (1) in the proof of estimation of consistency in Theorem 1, we can subsequently show that $|E_1| \leq O_p(1)$. A similar development to be above can be constructed to show that $|E_2| \leq O_p(1)$. It follows that $|D| \leq O_p(1)$, and thus by applying this result to (6) we obtain $N\{\mathrm{IC}_{\text{proxy}}(\alpha) - \mathrm{IC}_{\text{proxy}}(\alpha_0)\} > -2 \times O_p(1) + \log(n)d_f + 2nd_r$. Since the right hand side of the inequality tends to infinity as $n \to \infty$, it follows that for all overfitted models the difference $\{\mathrm{IC}_{\text{proxy}}(\alpha) - \mathrm{IC}_{\text{proxy}}(\alpha_0)\}$ is guaranteed to be positive with probability tending to one.

Finally, combining the derivations for the cases of overfitting and underfitting leads to the result as announced. $\square$

Proof of Theorem 2

We begin by proving the following statement linking the penalized PQL estimator to the unregularized PQL estimator:

$$\frac{1}{N}\ell_{\mathrm{PQL}}(\hat{\boldsymbol{\Psi}}_{\lambda_0}, \hat{\boldsymbol{b}}_{\lambda_0}) = \frac{1}{N}\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) + o_p\left(1\right), \tag{7}$$

where $(\hat{\boldsymbol{\Psi}}^T_{\lambda_0}, \hat{\boldsymbol{b}}^T_{\lambda_0})^T$ are the regularized PQL estimates obtained at the tuning parameter λ_0 that selects the true model, $(\tilde{\boldsymbol{\beta}}^T_{\alpha_0}, \tilde{\boldsymbol{b}}^T_{\alpha_0})^T$ are the unregularized PQL estimates obtained at the true model α_0. To prove (7), write

$$\frac{1}{N}\ell_{\mathrm{PQL}}(\hat{\boldsymbol{\Psi}}_{\lambda_0}, \hat{\boldsymbol{b}}_{\lambda_0}) = \frac{1}{N}\ell_{\mathrm{PQL}}(\hat{\boldsymbol{\Psi}}_{\alpha_0}, \hat{\boldsymbol{b}}_{\alpha_0}) = \frac{1}{N}\ell_1(\hat{\boldsymbol{\beta}}_{\alpha_0}, \hat{\boldsymbol{b}}_{\alpha_0}) - \frac{1}{2N}\sum_{i=1}^{n}\hat{\boldsymbol{b}}^T_{\alpha_0 i}\hat{\boldsymbol{D}}^{-}_{\alpha_0}\hat{\boldsymbol{b}}_{\alpha_0 i},$$

where $\hat{\boldsymbol{\Psi}}_{\alpha_0}$ consists of only the non-zero elements of $\hat{\boldsymbol{\Psi}}_{\lambda_0}$ and similarly for $\hat{\boldsymbol{b}}_{\alpha_0}$, remembering importantly that α_0 corresponds to the true model. From the estimation consistency in Theorem 1, we have $\hat{\boldsymbol{b}}^T_{\alpha_0 i}\hat{\boldsymbol{D}}^{-}_{\alpha_0}\hat{\boldsymbol{b}}_{\alpha_0 i} \xrightarrow{p} \boldsymbol{b}^T_{01i}\boldsymbol{D}^{-}_{01}\boldsymbol{b}_{01i}$ and therefore $(2N)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{b}}^T_{\alpha_0 i}\hat{\boldsymbol{D}}^{-}_{\alpha_0}\hat{\boldsymbol{b}}_{\alpha_0 i} = O_p(nN^{-1}) = o_p(1)$. For $\ell_1(\hat{\boldsymbol{\beta}}_{\alpha_0}, \hat{\boldsymbol{b}}_{\alpha_0})$, a Taylor expansion about the unregularized PQL estimator yields

$$\begin{aligned}\ell_1(\hat{\boldsymbol{\beta}}_{\alpha_0}, \hat{\boldsymbol{b}}_{\alpha_0}) &= \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) + \nabla\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0})^T\left(\hat{\boldsymbol{\beta}}^T_{\alpha_0} - \tilde{\boldsymbol{\beta}}^T_{\alpha_0}, \hat{\boldsymbol{b}}^T_{\alpha_0} - \tilde{\boldsymbol{b}}^T_{\alpha_0}\right) \\ &\quad - \frac{N}{2}\left(\hat{\boldsymbol{\beta}}^T_{\alpha_0} - \tilde{\boldsymbol{\beta}}^T_{\alpha_0}, \hat{\boldsymbol{b}}^T_{\alpha_0} - \tilde{\boldsymbol{b}}^T_{\alpha_0}\right)^T\left\{-\frac{1}{N}\nabla^2\ell_1(\bar{\boldsymbol{\beta}}_{\alpha_0}, \bar{\boldsymbol{b}}_{\alpha_0})\right\}\left(\hat{\boldsymbol{\beta}}^T_{\alpha_0} - \tilde{\boldsymbol{\beta}}^T_{\alpha_0}, \hat{\boldsymbol{b}}^T_{\alpha_0} - \tilde{\boldsymbol{b}}^T_{\alpha_0}\right) \\ &\triangleq T_1 + T_2 + T_3,\end{aligned} \tag{8}$$

where $(\bar{\boldsymbol{\beta}}^T_{\alpha_0}, \bar{\boldsymbol{b}}^T_{\alpha_0})^T$ lies on the line segment joining $(\hat{\boldsymbol{\beta}}^T_{\alpha_0}, \hat{\boldsymbol{b}}^T_{\alpha_0})^T$ and $(\tilde{\boldsymbol{\beta}}^T_{\alpha_0}, \tilde{\boldsymbol{b}}^T_{\alpha_0})^T$. Next, we recognize that by Theorem 1, the regularized PQL estimator satisfies $\hat{\boldsymbol{\beta}}_{\alpha_0} - \boldsymbol{\beta}_{01} = O_p(m_1^{-1/2})$ and $\hat{\boldsymbol{b}}_{\alpha_0 i} - \boldsymbol{b}_{01i} = O_p(m_1^{-1/2}); i = 1, \ldots, n$. Moreover, by equation (3), we know that $m_1^{1/2}$-consistency also holds for the unregularized PQL estimator $(\tilde{\boldsymbol{\beta}}^T_{\alpha 0}, \tilde{\boldsymbol{b}}^T_{\alpha 0})^T$. Combining these results, we obtain $\hat{\boldsymbol{\beta}}_{\alpha 0} - \tilde{\boldsymbol{\beta}}_{\alpha 0} = O_p(m_1^{-1/2})$ and, for all $i = 1, \ldots, n$, $\hat{\boldsymbol{b}}_{\alpha 0 i} - \tilde{\boldsymbol{b}}_{\alpha 0 i} = O_p(m_1^{-1/2})$. Furthermore, since $(\bar{\boldsymbol{\beta}}^T_{\alpha 0}, \bar{\boldsymbol{b}}^T_{\alpha 0})^T$ lies on the line segment joining $(\hat{\boldsymbol{\beta}}^T_{\alpha 0}, \hat{\boldsymbol{b}}^T_{\alpha 0})^T$ and $(\tilde{\boldsymbol{\beta}}^T_{\alpha 0}, \tilde{\boldsymbol{b}}^T_{\alpha 0})^T$, then $(\bar{\boldsymbol{\beta}}^T_{\alpha 0}, \bar{\boldsymbol{b}}^T_{\alpha 0})^T$ must lie within the $m_1^{1/2}$ neighborhood of $(\boldsymbol{\beta}^T_{01}, \boldsymbol{b}^T_{01})^T$. Therefore, there exists an $\varepsilon > 0$ satisfying $\|(\bar{\boldsymbol{\beta}}^T_{\alpha}, \bar{\boldsymbol{b}}^T_{\alpha})^T - (\boldsymbol{\beta}^T_{0}, \boldsymbol{b}^T_{0})^T\|_\infty \leq \varepsilon$ for n and all m_i large enough. Applying condition (C3), the Hessian matrix $\boldsymbol{H}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0}) = -N^{-1}\nabla^2 \ell_1(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})$ is invertible so we can apply the inequality

$$-\frac{N}{2}\left\{\left(\hat{\boldsymbol{\beta}}^T_{\alpha 0} - \tilde{\boldsymbol{\beta}}^T_{\alpha 0}, \hat{\boldsymbol{b}}^T_{\alpha 0} - \tilde{\boldsymbol{b}}^T_{\alpha 0}\right)^T - N^{-1}\boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0})\right\}^T \boldsymbol{H}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})$$
$$\times \left\{\left(\hat{\boldsymbol{\beta}}^T_{\alpha 0} - \tilde{\boldsymbol{\beta}}^T_{\alpha 0}, \hat{\boldsymbol{b}}^T_{\alpha 0} - \tilde{\boldsymbol{b}}^T_{\alpha 0}\right)^T - N^{-1}\boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0})\right\} \leq 0,$$

to obtain

$$T_2 + T_3 \leq \frac{1}{N}\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0})^T \boldsymbol{H}^{-1}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0}) \leq \frac{1}{N\tau_{\min}}\|\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0})\|^2,$$

where $\tau_{\min} > 0$ denotes the minimum eigenvalue of $\boldsymbol{H}(\bar{\boldsymbol{\beta}}_{\alpha 0}, \bar{\boldsymbol{b}}_{\alpha 0})$ by condition (C3) and is bounded away from zero.

By definition of the unregularized PQL estimator, $\partial \ell_1(\boldsymbol{\beta}_{\alpha 0}, \boldsymbol{b}_{\alpha 0})/\partial \boldsymbol{\beta}_{\alpha 0}\big|_{\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0}} = \mathbf{0}$ and $\partial \ell_1(\boldsymbol{\beta}_{\alpha 0}, \boldsymbol{b}_{\alpha 0})/\partial \boldsymbol{b}_{\alpha 0 i}\big|_{\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0}} = \tilde{\boldsymbol{D}}^{-}_{\alpha 0}\tilde{\boldsymbol{b}}_{\alpha 0 i}$ for all $i = 1, \ldots. n$. Furthermore, by the $m_1^{1/2}$-consistency of $(\tilde{\boldsymbol{\beta}}^T_{\alpha 0}, \tilde{\boldsymbol{b}}^T_{\alpha 0})^T$ as discussed earlier, we can show that $\tilde{\boldsymbol{D}}_{\alpha 0}$ is also a $m_1^{1/2}$-consistent estimator of the true random effects covariance matrix. and hence $\tilde{\boldsymbol{D}}^{-}_{\alpha 0}\tilde{\boldsymbol{b}}_{\alpha 0 i} = O_p(1)$. We thus conclude $T_2 + T_3 \leq (N\tau_{\min})^{-1}\|\nabla \ell_1(\tilde{\boldsymbol{\beta}}_{\alpha 0}, \tilde{\boldsymbol{b}}_{\alpha 0})\|^2 = O_p(nN^{-1}) = o_p(1)$, and equation (7) follows by applying this order result to (8).

We now utilize (7) to prove the theorem. First, note that by definition of the unregularized PQL estimator, $-2\ell_{\text{PQL}}(\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda) \geq -2\ell_{\text{PQL}}(\tilde{\boldsymbol{\Psi}}_\alpha, \tilde{\boldsymbol{b}}_\alpha) \geq -2\ell_1(\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_\alpha)$, where the second inequality follows since $\tilde{\boldsymbol{b}}_{\alpha i}^T \tilde{\boldsymbol{D}}_{\alpha}^{-} \tilde{\boldsymbol{b}}_{\alpha i} \geq 0$. Combining this with equation (7), we obtain

$$-\frac{2}{N}\ell_{\text{PQL}}(\hat{\boldsymbol{\Psi}}_{\lambda}, \hat{\boldsymbol{b}}_{\lambda}) + \frac{2}{N}\ell_{\text{PQL}}(\hat{\boldsymbol{\Psi}}_{\lambda_0}, \hat{\boldsymbol{b}}_{\lambda_0}) \geq \\ -\frac{2}{N}\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha}, \tilde{\boldsymbol{b}}_{\alpha}) + \frac{2}{N}\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) + o_p(1),$$

from which we obtain

$$\text{IC}(\lambda) - \text{IC}(\lambda_0) \geq \text{IC}_{\text{proxy}}(\alpha) - \text{IC}_{\text{proxy}}(\alpha_0) + o_p(1).$$

From Lemma 1, the right hand side of the above is asymptotically greater than zero for all $\alpha \neq \alpha_0$. It follows immediately that for any λ producing an overfitted or underfitted model, the difference $\text{IC}(\lambda) - \text{IC}(\lambda_0)$ is guaranteed to be asymptotically greater than zero. $\square$

4 Additional Simulation Results

Bernoulli Responses

We considered two additional methods of forward selection using other information criteria, specifically with: 1) a BIC with model complexity penalty depending on cluster size only

$$\text{BIC}_1(\alpha) = -2\ell(\tilde{\Psi}_{\alpha}) + \log(n)\dim(\tilde{\boldsymbol{\Psi}}_{\alpha}),$$

and 2) marginal corrected AIC (Vaida and Blanchard, 2005),

$$\text{AICc}(\alpha) = -2\ell(\tilde{\Psi}_\alpha) + 2\dim(\tilde{\boldsymbol{\Psi}}_\alpha)\frac{N}{N - \dim(\tilde{\boldsymbol{\Psi}}_\alpha) - 1}$$

Results show both these criteria performed worse than regularized PQL and the BIC(α) in Section 5.2 of the the main text.

Table 1: Additional results from simulation Setting 2 for Bernoulli GLMMs. The methods are: forward selection using $\text{BIC}_1(\alpha)$ (FS-BIC_1), and forward selection using marginal corrected AIC, AICc(α), (FS-AICc). Performance was assessed in terms of the percentage of datasets with correctly chosen overall models (%C), fixed effects (%CF), and random effects (%CR). The mean computation time in seconds (standard deviation in parenthesis) was also recorded.

(n, m)	Method	%C	% CF	% CR	Comp. time
(50,10)	FS-BIC_1	28	78	35	220(95)
	FS-AICc	28	39	51	332(143)
(50,20)	FS-BIC_1	69	76	92	1917(405)
	FS-AICc	27	28	75	2260(429)
(100,10)	FS-BIC_1	51	80	61	1993(440)
	FS-AICc	22	28	68	2724(525)
(100,20)	FS-BIC_1	82	82	100	6711(1076)
	FS-AICc	32	35	78	7527(1279)

Poisson Response

We simulated count datasets from a Poisson GLMM. We considered a model with $p_f = p_r = 9$ covariates including an intercept term, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ generated in the same manner as the Bernoulli GLMMs in Section 5.2 of the main text. The vector of true effects was also set to the same one used in the Bernoulli GLMMs simulation, while the true random effect covariance matrix was a 9×9 matrix with the top left 3×3 submatrix set to

$$\begin{pmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{pmatrix}$$

and all remaining elements equal to zero. This setting presents a rather challenging selection problem for Poisson GLMMs; the counts are heavily right skewed, with an average of 40% of the entries in a simulated response matrix equal to zero.

As in Section 5.2 of the main text, we compared regularized PQL with the following methods: 1) the M-ALASSO penalty of Bondell et al. (2010), with estimation performed using the Monte-Carlo EM algorithm and tuning parameter chosen using the BIC, 2) the adaptive lasso penalty of Ibrahim et al. (2011), with estimation performed using the Monte-Carlo EM algorithm and tuning parameters chosen using IC_Q, 3) `glmmlasso` assuming the true random effects structure us known, 4) two-stage, forward selection approaches using $\text{BIC}(\alpha) = -2\ell(\tilde{\Psi}_\alpha) + \log(N)\dim(\tilde{\boldsymbol{\Psi}}_\alpha)$ for submodel α, as well as $\text{BIC}_1(\alpha)$ and marginal corrected AIC, denoted as $\text{AICc}(\alpha)$, shown in the "Bernoulli Responses" section above.

When implementing the `glmmLasso` method for this setting, we limited the maximum number of iterations for each fit to 200 (as opposed to the default setting of 1000 iterations that we used in Bernoulli responses in Section 5.2 of the main text) for computational reasons. For instance, at $(n, m) = (50, 20)$, `glmmLasso` took in excess of one hour to fit the full regularization path for λ, while regularized PQL fitted a path with the same number of λ values to perform joint selection in approximately three minutes.

Regularized PQL performed well in selecting the random effects, but was poorer in selecting the fixed effects particularly at $n = 50$ (Table 2). This contrasted to Setting 2 with binary responses, where regularized PQL performed better at selecting the fixed than the random effects. Forward selection using both versions of BIC performed best overall, although its computation time scaled poorly with increasing n and m. Regularized PQL outperformed the all three penalized likelihood methods i.e., M-ALASSO, the penalty of Ibrahim et al. (2011) and `glmmLasso`, while also taking substantially less time than all other methods for joint selection. It is also interesting to note though that the penalty of Ibrahim et al. (2011) took considerably less time to perform joint selection than it did in the Bernoulli response setting.

In contrast to the Bernoulli response case, the penalty of Ibrahim et al. (2011) performed reasonably in this setting, even though further investigation showed that IC_Q continued to behave rather erratically, with the loss function component of the criterion varying non-monotonically with model complexity. We hypothesize that one reason why IC_Q may have performed reasonably here is because of Poisson responses contain substantially more information than Bernoulli responses (noting that IC_Q also performed reasonably well for the Gaussian responses simulations in Ibrahim et al., 2011). Finally, similar to the Bernoulli response setting in the main text, we see that the hybrid estimator produces considerably less biased but more variable estimates than the regularized PQL estimates, although the ratios here are much less extreme (far away from 1) than in the Bernoulli setting.

Table 2: Results from simulation Setting 3 for Poisson GLMMs. The methods are: regularized PQL (rPQL), M-ALASSO (Bondell et al., 2010), I-ALASSO (Ibrahim et al., 2011), glmmLasso (Groll and Tutz, 2014), and forward selection using BIC(α) (FS-BIC), $\text{BIC}_1(\alpha)$ (FS-BIC_1) and marginal corrected AIC, AICc(α), (FS-AICc). Performance was assessed in terms of the percentage of datasets with correctly chosen overall models (%C), fixed effects (%CF), and random effects (%CR), as well as the mean ratio in predicted log-likelihood (standard deviation in parenthesis), the ratios of mean absolute bias (Bias) and total variance (Var) of the estimates, mean squared prediction error (PSE), and predicted log-likelihood (PL). Finally, the mean computation time for each method was also recorded, with standard deviations in parentheses. glmmLasso was implemented assuming the true random effects component was known.

(n, m)	Method	%C	% CF	% CR	Comp. time	Bias/Var/PSE/PL
	rPQL	55	55	100	184(34)	0.75/1.36/0.86/0.57
	M-ALASSO	14	61	16	$\approx 20^4$	-
	I-ALASSO	24	42	57	2737(401)	-
(50, 10)	glmmLasso	-	12	-	1625(71)	-
	FS-BIC	91	91	100	273(62)	-
	FS-BIC_1	74	74	100	277(63)	-
	FS-AICc	39	40	85	191(61)	-
	rPQL	63	65	99	283(59)	0.87/1.09/0.89/0.75
	M-ALASSO	47	72	65	$\approx 30^4$	-
	I-ALASSO	38	51	81	2751(524)	-
(50, 20)	glmmLasso	-	47	-	2273(208)	-
	FS-BIC	94	94	100	1164(263)	-
	FS-BIC_1	82	82	100	1190(270)	-
	FS-AICc	38	41	72	1126(227)	-
	rPQL	69	69	100	378(51)	0.73/1.30/0.95/0.60
	M-ALASSO	26	71	30	$\approx 40^4$	-
	I-ALASSO	46	57	75	4577(622)	-
(100, 10)	glmmLasso	-	44	-	5461(313)	-
	FS-BIC	95	95	100	1442(331)	-
	FS-BIC_1	88	88	100	1453(329)	-
	FS-AICc	46	46	78	1663(349)	-
	rPQL	86	86	100	554(67)	0.81/1.05/0.96/0.80
	M-ALASSO	70	72	94	$\approx 50^4$	-
	I-ALASSO	56	70	83	4626(677)	-
(100, 20)	glmmLasso	-	74	-	8328(544)	-
	FS-BIC	94	94	100	2962(441)	-
	FS-BIC_1	89	89	100	2999(450)	-
	FS-AICc	44	45	80	3369(626)	-

Binomial Responses with m growing with n

We simulated Bernoulli datasets in the same manner as in Section 5.2 of the main text, except that $m = \lceil 6n^{1/2} \rceil$ where $\lceil \cdot \rceil$ denotes the ceiling function. In other words, the response matrices were relatively wide. Compared to the case where m was small relative to n (see Table 2), regularized PQL performed very well in in selecting both the fixed and random effects (Table 3). Furthermore, the mean ratio of predicted log-likelihoods was very close to one, indicating that there was little difference between the hybrid and regularized PQL estimates. This is compatible with the estimation consistency as provided in Theorem 1.

Finally, the difference in computation times between the two methods was significantly greater than in Section 5.2.

Table 3: Results from simulation with Bernoulli GLMMs with wide response matrices. Performance was assessed in terms of the percentage of datasets with correctly chosen overall models (%C), fixed effects (%CF), and random effects (%CR), as well as the mean ratio in predicted log-likelihood (standard deviation in parenthesis). The mean computation time in seconds (standard deviation in parenthesis) was also recorded.

n	m	%C	% CF	% CR	Comp. time	Ratio
50	43	93	96	98	593(50)	0.985(0.019)
100	60	99	99	100	1081(85)	0.998(0.009)

We also performed a simulation with Poisson GLMMs using the same setup presented in the *Poisson GLMMs* simulation setting but also with $m = \lceil 6n^{1/2} \rceil$, i.e., wide response matrices. The overall results were very similar to those seen above for Bernoulli GLMM, and may be obtained upon request from the corresponding author.

References

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. Wiley Series in Probability and Statistics. Wiley.

Ferentinos, K. (1982). On Tchebycheff's type inequalities. *Trabajos de Estadstica e Investigacin Operativa*, 33:125–132.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by ℓ_1–penalized estimation. *Statistics and Computing*, 24:137–154.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503.

Jiang, J., Jia, H., and Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica*, 11:97–120.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *Journal of Statistical Planning and Inference*, 137:1787–1804.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83:447–452.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1–26.