HIERARCHICAL SELECTION OF FIXED AND RANDOM EFFECTS IN GENERALIZED LINEAR MIXED MODELS

The Australian National University and The University of Sydney

Supplementary Material

S1 Proofs

S1.1 Proof of Theorem 1

For simplicity, the derivations below are performed assuming ϕ is known, as in binomial and Poisson GLMMs where $\phi = 1$. The extension to the case of unknown ϕ is straightforward, because it does not appear in the CREPE penalty and is of order $\hat{\phi} = O_p(1)$ for all n and λ. Also, we focus on the case with the CREPE penalty defined in equation (1) of the main text, as opposed to the modification of it to account for a penalized random intercept but unpenalized fixed intercept. The developments below can be straightforwardly extended to the case where the CREPE penalty is defined as $n\lambda(\tilde{v}_1\|\boldsymbol{\gamma}_1\|)^{1/2} + n\lambda \sum_{k=2}^{p} \tilde{w}_k(\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_l\|\boldsymbol{\gamma}_k\|)^{1/2}$.

We first prove estimation consistency. To begin, rewrite equation (1) in

the main text as $\ell_{pen}(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - n\lambda \sum_{k \in \alpha_f} \tilde{w}_k |\beta_k| - n\lambda \sum_{l \in \alpha_c} \tilde{w}_l (\beta_l^2 + \tilde{v}_l \|\boldsymbol{\gamma}_l\|)^{1/2}$. Let $r_n = \sqrt{p_f/n}$, and $D(\boldsymbol{u}) = \ell_{pen}(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}) - \ell_{pen}(\boldsymbol{\Psi}_0)$. We want to show that for any given $\varepsilon > 0$, there exists a constant C such that for sufficiently large n,

$$P\left(\sup_{\|\boldsymbol{u}\| = C} D(\boldsymbol{u}) < 0 \right) \geq 1 - \varepsilon. \tag{S1.1}$$

If the above holds, then it guarantees that there exists a local maximizer $\ell_{pen}(\boldsymbol{\Psi})$, denoted here as $\hat{\boldsymbol{\Psi}}$ of, such that $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\| = O_p\left(\sqrt{p_f/n}\right)$ (see Fan and Peng, 2004). To prove (S1.1), first note that

$$\begin{aligned} D(\boldsymbol{u}) \leq & \left\{ \ell(\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}) - \ell(\boldsymbol{\Psi}_0) \right\} - n\lambda \sum_{k \in \alpha_{0f}} \tilde{w}_k \left(|\beta_{0k} + r_n u_k| - |\beta_{0k}| \right) \\ & - n\lambda \sum_{l \in \alpha_{0c}} \tilde{w}_l \left[\left\{ (\beta_{0l} + r_n u_{l1})^2 + \tilde{v}_l \|\boldsymbol{\gamma}_{0l} + r_n \boldsymbol{u}_{l2}\| \right\}^{1/2} - \left(\beta_{0l}^2 + \tilde{v}_l \|\boldsymbol{\gamma}_{0l}\| \right)^{1/2} \right] \\ & \triangleq L_1 - L_2 - L_3, \end{aligned}$$

where $\alpha_{0f} = \{k \in \alpha_f : \beta_{0k} \neq 0\}$, $\alpha_{0c} = \{l \in \alpha_c : \beta_{0l} \neq 0\}$, and $\boldsymbol{u} = (u_1, \ldots, u_{p_f}, u_{11}, \boldsymbol{u}_{12}, u_{21}, \boldsymbol{u}_{22}, \ldots, u_{p_c 1}, \boldsymbol{u}_{p_c 2})$. Note that for elements in α_{0c}, the quantity $\|\boldsymbol{\gamma}_{0l}\|$ may or may not be equal to zero. That is, the subsets α_{0f} and α_{0c} are obtained by omitting truly zero fixed effects and truly zero composite effects, respectively. Put another way, the inequality in the first

line of the above expression comes from recognizing that: 1) for all $k \in \alpha_f$ where $\beta_{0k} = 0$, it holds that $(|\beta_{0k} + r_n u_k| - |\beta_{0k}|) \geq 0$, and 2) for all $l \in \alpha_c$ where $\beta_{0k} = \|\boldsymbol{\gamma}_{0l}\| = 0$, it holds that $\left[\{(\beta_{0l} + r_n u_{l1})^2 + \tilde{v}_l \|\boldsymbol{\gamma}_{0l} + r_n \boldsymbol{u}_{l2}\|\}^{1/2} - (\beta_{0l}^2 + \tilde{v}_l \|\boldsymbol{\gamma}_{0l}\|)^{1/2}\right] > 0$.

For term L_1, a Taylor expansion can be used to obtain

$$L_1 = r_n \boldsymbol{u}^T \nabla \ell(\boldsymbol{\Psi}_0) - \frac{1}{2} n r_n^2 \boldsymbol{u}^T \left(-\frac{1}{n} \nabla^2 \ell(\bar{\boldsymbol{\Psi}}) \right) \boldsymbol{u},$$

where $\bar{\boldsymbol{\Psi}}$ lies on the line segment joining $\boldsymbol{\Psi}_0$ and $\boldsymbol{\Psi}_0 + r_n \boldsymbol{u}$. A standard argument using Chebychev's inequality can be used to show that $\nabla \ell(\boldsymbol{\Psi}_0) = O_p(\sqrt{n p_f})$ (see Fan and Peng, 2004), from which we obtain $r_n \boldsymbol{u}^T \nabla \ell(\boldsymbol{\Psi}_0) = O_p(n r_n^2)$. Using Conditions (C1)-(C2) and the Cauchy-Schwarz inequality, we have that for sufficiently large n,

$$L_1 \leq r_n \boldsymbol{u}^T \nabla \ell(\boldsymbol{\Psi}_0) - \frac{1}{2} n r_n^2 \|\boldsymbol{u}\|^2 (1 - \epsilon) c_1, \tag{S1.2}$$

where ϵ is different to the one in (S1.1). Next, by the Cauchy-Schwarz inequality and condition (C5), we have $L_2 = n\lambda \sum_{k \in \alpha_{0f}} \tilde{w}_k r_n u_k \text{sgn}(\beta_{0k}) \leq O_p(n\lambda r_n \sqrt{p_{0f}}) = o_p(n r_n^2)$ by Condition (C6a), where $\text{sgn}(\cdot)$ denotes the sign function. Turning to term L_3, note that for n large enough, $(\beta_{0l} + r_n u_{l1})^2 > \beta_{0l}^2 - 2 r_n |u_{l1} \beta_{0l}|$. Moreover for n large enough, we have $\|\boldsymbol{\gamma}_{0l} + r_n \boldsymbol{u}_{l2}\| \geq$

$\|\boldsymbol{\gamma}_{0l}\| - 2r_u\|\boldsymbol{u}_{l2}\|$. Hence for sufficiently large n,

$$\begin{aligned} L_3 &\geq n\lambda \sum_{l\in\alpha_{0c}} \tilde{w}_l \left\{ \left(\beta_{0l}^2 - 2r_n|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{\gamma}_{0l}\| - 2r_u\tilde{v}_l\|\boldsymbol{u}_{l2}\|\right)^{1/2} - \left(\beta_{0l}^2 + \tilde{v}_l\|\boldsymbol{\gamma}_{0l}\|\right)^{1/2} \right\} \\ &= n\lambda \sum_{l\in\alpha_{0c}} \tilde{w}_l\xi_l^0 \left\{ \left(1 - \frac{2r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)}{(\xi_l^0)^2}\right)^{1/2} - 1 \right\}, \end{aligned}$$

where $\xi_l^0 = (\beta_{0l}^2 + \tilde{v}_l\|\boldsymbol{\gamma}_{0l}\|)^{1/2}$. Observe that by Condition (C4),

$$\begin{aligned} \frac{r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)}{(\xi_l^0)^2} &\leq \frac{r_nC_1(|u_{l1}| + \|\boldsymbol{u}_{l2}\|}{\min_{l\in\alpha_0}\{\beta_{0l}^2\} + \min_{l\in\alpha_0}\{\|\boldsymbol{\gamma}_{0l}\|\}} \\ &\leq \frac{r_nC_1(|u_{l1}| + \|\boldsymbol{u}_{l2}\|)}{c_2} \to 0, \end{aligned}$$

where $C_1 > 0$ is a sufficiently large constant and $\alpha_0 = \alpha_{0f} \cup \alpha_{0c}$. Using this result, we can apply a Taylor expansion $\sqrt{1-x} = 1 - (1/2)x + O_p(x^2)$, with $x = r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)(\xi_l^0)^{-2}$, to show that for n large enough,

$$\begin{aligned} L_3 &\geq -n\lambda \sum_{l\in\alpha_{0c}} \tilde{w}_l\xi_l^0 \left(\frac{r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)}{(\xi_l^0)^2}\right) \{1 + o_p(1)\} \\ &= -n\lambda \sum_{l\in\alpha_{0c}} \tilde{w}_l \left(\frac{r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)}{\xi_l^0}\right) \{1 + o_p(1)\}, \end{aligned}$$

where the $\{1 + o_p(1)\}$ follows since $x^2 = o_p(x)$ when $x = r_n(|u_{l1}\beta_{0l}| +$

$\tilde{v}_l\|\boldsymbol{u}_{l2}\|)(\xi_l^0)^{-2}$. Using condition (C5), we have

$$\begin{aligned}
-L_3 &\leq n\lambda \sum_{l\in\alpha_{0c}} \tilde{w}_l \left(\frac{r_n(|u_{l1}\beta_{0l}| + \tilde{v}_l\|\boldsymbol{u}_{l2}\|)}{\xi_l^0}\right)\{1+o_p(1)\} \\
&\leq \frac{n\lambda r_n C_2}{\sqrt{c_2}}\{1+o_p(1)\} \quad \text{by Condition (C4)} \\
&\leq O_p(n\lambda r_n) = o_p(nr_n^2) \quad \text{by Condition (C6a)},
\end{aligned}$$

for some sufficiently large constant $C_2 > 0$. Combining all the results above, we have that for sufficiently large $\|\boldsymbol{u}\| = C$, all the terms in $D(\boldsymbol{u})$ are dominated by the second term on the right hand side of (S1.2), which is negative. The statement in equation (S1.1) and the desired estimation consistency follows.

We now prove selection consistency. This will be done by considering three cases, where in each case it is demonstrated that if the true parameter is equal to zero, then with probability tending to one the corresponding CREPE estimate (which is estimation consistent from the proof above) must also equal zero.

First, suppose that for some $k \in \alpha_f$, we have $\beta_{0k} = 0$ but $\hat{\beta}_k \neq 0$. By the Karush-Kuhn-Tucker (KKT) optimality conditions,

$$0 = \left.\frac{\partial \ell_{pen}(\boldsymbol{\Psi})}{\partial \beta_k}\right|_{\hat{\boldsymbol{\Psi}}} = \left.\frac{\partial \ell(\boldsymbol{\Psi})}{\partial \beta_k}\right|_{\hat{\boldsymbol{\Psi}}} - n\lambda\tilde{w}_k \text{sgn}(\hat{\beta}_k). \tag{S1.3}$$

Using a Taylor expansion and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left.\frac{\partial \ell(\boldsymbol{\Psi})}{\partial \beta_k}\right|_{\hat{\boldsymbol{\Psi}}} &= \left.\frac{\partial \ell(\boldsymbol{\Psi})}{\partial \beta_k}\right|_{\boldsymbol{\Psi}_0} + \sum_{r \in \boldsymbol{\Psi}} \left.\frac{\partial^2 \ell(\boldsymbol{\Psi})}{\partial \beta_k \partial \Psi_r}\right|_{\bar{\boldsymbol{\Psi}}} (\hat{\Psi}_r - \Psi_{0r}) \\ &\leq \left.\frac{\partial \ell(\boldsymbol{\Psi})}{\partial \beta_k}\right|_{\boldsymbol{\Psi}_0} + n\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\| \left(\sum_{r \in \boldsymbol{\Psi}} \left(\frac{1}{n} \left.\frac{\partial^2 \ell(\boldsymbol{\Psi})}{\partial \beta_k \partial \Psi_r}\right|_{\bar{\boldsymbol{\Psi}}} \right)^2 \right)^{1/2} \\ &\triangleq M_1 + M_2.\end{aligned}$$

From the proof of estimation consistency above, we have $M_1 = O_p(\sqrt{np_f})$. Furthermore, by Conditions (C1)-(C2) and the estimation consistency of $\hat{\boldsymbol{\Psi}}$, we have that for sufficiently large n, $M_2 = O_p(n\sqrt{p_f/n}) = O_p(\sqrt{np_f})$, from which it follows that the first term on the right hand side of (S1.3) is $O_p(\sqrt{np_f})$. On the other hand, by Condition (C5), we have $\tilde{w}_k = O_p\{(n/p_f)^{\nu/2}\}$. It follows that for $\nu \geq 1$, $n\lambda\tilde{w}_k/\sqrt{np_f} = O\left\{(n/p_f)^{(\nu+1)/2}\right\} \geq O\left\{(n/p_f)^{(\nu+3)/4}\right\} \to \infty$ by Condition (C6b). It follows from the above that the second term on the right hand side of equation (S1.3) asymptotically dominates the first term. With probability tending to one, the right hand side of equation (S1.3) cannot equal zero. We therefore have a contradiction, from which we conclude that for all $k \in \alpha_f$ with $\beta_{0k} = 0$, $P(\hat{\beta}_k = 0) \to 1$.

Suppose now that for some $l \in \alpha_c$, we have $\beta_{0l} = 0$ but $\hat{\beta}_l \neq 0$. Note that by the definition of a truly zero composite effect, $\beta_{0l} = 0$ implies

$\|\boldsymbol{\gamma}_{0l}\| = 0$ for all $l \in \alpha_c$. Now, by the KKT optimality conditions,

$$0 = \frac{1}{\sqrt{np_f}} \frac{\partial \ell_{pen}(\boldsymbol{\Psi})}{\partial \beta_l}\bigg|_{\hat{\boldsymbol{\Psi}}} = \frac{1}{\sqrt{np_f}} \frac{\partial \ell(\boldsymbol{\Psi})}{\partial \beta_l}\bigg|_{\hat{\boldsymbol{\Psi}}} - \frac{1}{\sqrt{np_f}} \frac{n\lambda \tilde{w}_l \hat{\beta}_l}{\left(\hat{\beta}_l^2 + \tilde{v}_l \|\hat{\boldsymbol{\gamma}}_l\|\right)^{1/2}}. \tag{S1.4}$$

We have $(1/\sqrt{np_f})\partial \ell(\boldsymbol{\Psi})/\partial \beta_l|_{\hat{\boldsymbol{\Psi}}} = O_p(1)$. By definition, $\hat{\beta}_l/(\hat{\beta}_l^2 + \tilde{v}_l\|\hat{\boldsymbol{\gamma}}_l\|)^{1/2} \in [-1, 1]$ for all n, and so we need only consider the order of $n\lambda\tilde{w}_l/\sqrt{np_f}$. As in (S1.3), we have that the second term on the right hand side of (S1.4) asymptotically dominates the first term, and therefore with probability tending to one, equation (S1.4) cannot equal zero. A contraction is thus obtained, from which it follows that for all $l \in \alpha_{0c}$, $P(\hat{\beta}_l = 0) \to 1$.

We turn to the third part of the proof of selection consistency. Suppose for some $l \in \alpha_c$, it holds that $\|\boldsymbol{\gamma}_{0l}\| = 0$ but $\|\hat{\boldsymbol{\gamma}}_l\| \neq 0$. By the design of the CREPE penalty, this implies $\hat{\beta}_l \neq 0$. From the KKT optimality conditions, we have for all $m = 1, \ldots, p_c$,

$$0 = \frac{1}{\sqrt{np_f}} \frac{\partial \ell_{pen}(\boldsymbol{\Psi})}{\partial \gamma_{lm}}\bigg|_{\hat{\boldsymbol{\Psi}}} = \frac{1}{\sqrt{np_f}} \frac{\partial \ell(\boldsymbol{\Psi})}{\partial \gamma_{lm}}\bigg|_{\hat{\boldsymbol{\Psi}}} - \frac{1}{\sqrt{np_f}} \frac{n\lambda \tilde{w}_l}{2\left(\hat{\beta}_l^2 + \tilde{v}_l \|\hat{\boldsymbol{\gamma}}_l\|\right)^{1/2}} \frac{\tilde{v}_l \hat{\gamma}_{lm}}{\|\hat{\boldsymbol{\gamma}}_l\|}. \tag{S1.5}$$

Similar to above, we have that $(1/\sqrt{np_f})\partial \ell(\boldsymbol{\Psi})/\partial \gamma_{lm}|_{\hat{\boldsymbol{\Psi}}} = O_p(1)$. By definition $\hat{\gamma}_{lm}/\|\hat{\boldsymbol{\gamma}}_l\| \in [-1, 1]$ for all n, and so we need only consider the order

of the quantity

$T_1 = (1/\sqrt{np_f})n\lambda\tilde{w}_l\tilde{v}_l \Big/ (\hat{\beta}_l^2 + \tilde{v}_l\|\hat{\boldsymbol{\gamma}}_l\|)^{1/2}$. We now consider two cases. First, suppose $\beta_{0l} \neq 0$. That is, the covariate enters the model as a composite effect, but it is in fact an important fixed effect only. Then by condition (C5), we have $\tilde{w}_l = O_p(1)$ and $\tilde{v}_l = O_p\{(n/p_f)^{\nu/2}\}$. Furthermore, by the estimation consistency of $\hat{\boldsymbol{\Psi}}$, we have $\hat{\beta}_l^2 = O_p(1)$ since $\beta_{0l} \neq 0$, and $\|\hat{\boldsymbol{\gamma}}_l\| = O_p\{(p_f/n)^{1/2}\}$ since $\|\boldsymbol{\gamma}_{0l}\| = 0$. It follows that $\tilde{v}_l\|\hat{\boldsymbol{\gamma}}_l\| = O_p\{(n/p_f)^{(\nu-1)/2}\}$ Given $\nu \geq 1$, then $(\hat{\beta}_l^2 + \tilde{v}_l\|\hat{\boldsymbol{\gamma}}_l\|)^{-1/2}$ has a lower bound of order $O_p(1)$. We therefore obtain $T_1 = O_p\left(\lambda\,(n/p_f)^{(\nu+3)/4}\right) \to \infty$ by Condition (C6b).

Suppose now $\beta_{0l} = 0$. That is, the covariate enters the model as a composite effect, but it is in fact a truly zero composite effect. Then $\tilde{w}_l = O_p\{(n/p_f)^{\nu/2}\}$ by condition (C5), and by the estimation consistency of $\hat{\boldsymbol{\Psi}}$ we have that $(\hat{\beta}_l^2+\tilde{v}_l\|\hat{\boldsymbol{\gamma}}_l\|)^{-1/2}$ has a lower bound of order $O_p\{(n/p_f)^{(\nu-1)/4}\}$. We thus obtain $T_1 = O_p\left\{\lambda\,(n/p_f)^{3(\nu+1)/4}\right\} \to \infty$ by condition (C6b). Combining the two cases above, we have that the right hand side of equation (S1.5) cannot equal zero with probability tending to one, and a contradiction is achieved. It follows that for any $l \in \alpha_{0c}$, if $\|\boldsymbol{\gamma}_{0l}\| = 0$ then $P(\|\hat{\boldsymbol{\gamma}}_l\| \neq 0) \to 1$.

Combining all the three proofs by contradiction leads to the result $P(\hat{\boldsymbol{\Psi}}_2 = \mathbf{0}) \to 1$, as desired. □

Proof of Theorem 2

From Theorem 1, we know that $\hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\Psi}}_1, \hat{\boldsymbol{\Psi}}_2 = \mathbf{0})$ is a $\sqrt{n/p_f}$-consistent local maximizer of the penalized log-likelihood in (1) in the main text. Thus in a slight abuse of notation, let $\ell_{pen}(\boldsymbol{\Psi}_1) = \ell_{pen}(\boldsymbol{\Psi}_1, \mathbf{0})$ and $\ell(\boldsymbol{\Psi}_1) = \ell(\boldsymbol{\Psi}_1, \mathbf{0})$. Letting $\rho_\lambda(\boldsymbol{\Psi}) = n\lambda \sum_{k=1}^{p} \tilde{w}_k(\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_k\|\boldsymbol{\gamma}_k\|)^{1/2}$ denote the CREPE penalty, we know that $\hat{\boldsymbol{\Psi}}_1$ must satisfy $\nabla\ell_{pen}(\hat{\boldsymbol{\Psi}}_1) = 0$. Therefore, we can construct the following Taylor expansion.

$$
\begin{aligned}
\mathbf{0} &= \frac{1}{\sqrt{n}}\nabla\ell_{pen}(\hat{\boldsymbol{\Psi}}_1) \\
&= \frac{1}{\sqrt{n}}\nabla\ell(\hat{\boldsymbol{\Psi}}_1) - \frac{1}{\sqrt{n}}\nabla\rho_\lambda(\hat{\boldsymbol{\Psi}}_1) \\
&= \frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\Psi}_{01}) + \frac{1}{\sqrt{n}}\nabla^2\ell(\boldsymbol{\Psi}_{01})(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{1}{2\sqrt{n}}R(\bar{\boldsymbol{\Psi}}_1) \\
&\quad - \frac{1}{\sqrt{n}}\nabla\rho_\lambda(\boldsymbol{\Psi}_{01}) - \frac{1}{\sqrt{n}}\nabla^2\rho_\lambda(\check{\boldsymbol{\Psi}}_1)(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) \\
&\triangleq T_1 + T_2 + T_3 + T_4 + T_5
\end{aligned}
$$

where $R(\bar{\boldsymbol{\Psi}}_1)$ is a vector remainder term with elements

$$
[R(\bar{\boldsymbol{\Psi}}_1)]_r = \sum_{s,t\in\boldsymbol{\Psi}_{01}} \frac{\partial^3\ell(\boldsymbol{\Psi}_1)}{\partial\Psi_r\partial\Psi_s\partial\Psi_t}\bigg|_{\bar{\boldsymbol{\Psi}}} (\hat{\Psi}_{1s} - \Psi_{01s})(\hat{\Psi}_{1t} - \Psi_{01t}),
$$

and the quantities $\check{\boldsymbol{\Psi}}_1$ and $\bar{\boldsymbol{\Psi}}_1$ both lie on the line segment joining $\hat{\boldsymbol{\Psi}}_1$ and $\boldsymbol{\Psi}_{01}$, and are not necessarily equal. We now consider the order of the terms

T_3 to T_5. Starting with T_3, by the estimation consistency of $\hat{\boldsymbol{\Psi}}$ it holds that $\|\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\|^2 = O_p(p_f/n)$. Therefore applying the Cauchy-Schwarz inequality and condition (C3), we have

$$\begin{aligned}\|T_3\| &\leq \frac{1}{2\sqrt{n}}\|\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\|^2 \times n \times \left(\sum_{r,s,t\in\boldsymbol{\Psi}_{01}} U_{rst}^2(\bar{\boldsymbol{\Psi}}_1)\right)^{1/2} \\ &\leq O_p\left(\sqrt{n} \times \frac{p_f}{n} \times p_{0f}^{3/2}\right) = o_p\left(\sqrt{\frac{p_{0f}}{p_f}}\right) = o_p(1),\end{aligned}$$

where the multiplier of n in the right hand side of the first line comes from noting that we have n independent clusters contributing to the marginal log-likelihood, $\ell(\boldsymbol{\Psi}_1) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\Psi}_1)$. Turning to T_4, observe that

$$\nabla\rho_\lambda(\boldsymbol{\Psi}_{01}) = \left(\frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\beta_k}\bigg|_{\boldsymbol{\Psi}_{01}}, \frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\beta_l}\bigg|_{\boldsymbol{\Psi}_{01}}, \frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\gamma_{lm}}\bigg|_{\boldsymbol{\Psi}_{01}}\right),$$

for $k \in \alpha_{0f}$ and $l, m \in \alpha_{0c}$, where

$$\begin{aligned}\frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\beta_k}\bigg|_{\boldsymbol{\Psi}_{01}} &= n\lambda\tilde{w}_k \mathrm{sgn}(\beta_{0k}), \\ \left|\frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\beta_l}\bigg|_{\boldsymbol{\Psi}_{01}}\right| &= \frac{n\lambda\tilde{w}_l|\beta_{0l}|}{(\beta_{0l}^2 + \tilde{v}_l\|\boldsymbol{\gamma}_{0l}\|)^{1/2}} \leq n\lambda\tilde{w}_l, \\ \left|\frac{\partial p_\lambda(\boldsymbol{\Psi}_1)}{\partial\gamma_{lm}}\bigg|_{\boldsymbol{\Psi}_{01}}\right| &= \frac{n\lambda\tilde{w}_l|\beta_{0l}|}{(\beta_{0l}^2 + \tilde{v}_l\|\boldsymbol{\gamma}_{0l}\|)^{1/2}} \frac{\tilde{v}_l|\gamma_{0lm}|}{\|\boldsymbol{\gamma}_{0l}\|} \leq n\lambda\tilde{w}_l\tilde{v}_l.\end{aligned}$$

By condition (C6a) then, it is straightforward to show that

$$\|T_4\| = \frac{1}{\sqrt{n}}\|\nabla\rho_\lambda(\boldsymbol{\Psi}_{01})\| \leq O_p\left(\lambda\sqrt{p_{0f}n}\right) = o_p(1).$$

Similarly, it can be shown that $\|T_5\| = o_p(1)$. Combining the above results, we have

$$\mathbf{0} = \frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\Psi}_{01}) + \frac{1}{\sqrt{n}}(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01})\nabla^2\ell(\boldsymbol{\Psi}_{01}) + o_p(1) \tag{S1.6}$$

We next prove the following result relating the expected and observed Fisher information matrices of $\boldsymbol{\Psi}_{01}$,

$$\left\|\frac{1}{n}\nabla^2\ell(\boldsymbol{\Psi}_{01}) + \mathcal{I}(\boldsymbol{\Psi}_{01})\right\| = o_p\left(\frac{1}{\sqrt{p_f}}\right), \tag{S1.7}$$

where $\mathcal{I}(\boldsymbol{\Psi}_{01})$ is the block of the expected Fisher information matrix involving only $\boldsymbol{\Psi}_{01}$. The above can be shown by applying Markov's inequality,

$$\begin{aligned}
P\left(\left\|\frac{1}{n}\nabla^2\ell(\boldsymbol{\Psi}_{01}) + \mathcal{I}(\boldsymbol{\Psi}_{01})\right\| > \frac{1}{\sqrt{p_f}}\right) &\leq p_f\mathrm{E}\left(\left\|\frac{1}{n}\nabla^2\ell(\boldsymbol{\Psi}_{01}) + \mathcal{I}(\boldsymbol{\Psi}_{01})\right\|^2\right) \\
&\leq \frac{p_f}{n^2}\mathrm{E}\left[\sum_{r,s\in\boldsymbol{\Psi}_{01}}\left\{\left.\frac{\partial^2\ell(\boldsymbol{\Psi})}{\partial\Psi_r\partial\Psi_s}\right|_{\boldsymbol{\Psi}_{01}} - \mathrm{E}\left(\left.\frac{\partial^2\ell(\boldsymbol{\Psi})}{\partial\Psi_r\partial\Psi_s}\right|_{\boldsymbol{\Psi}_{01}}\right)\right\}^2\right] \\
&\leq O_p\left(\frac{p_f}{n}p_{0f}^2\right) = o_p(1)
\end{aligned}$$

where the second line follows from the independence of the clusters $i = 1, \ldots, n$. Writing $n^{-1/2}(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01})\nabla^2\ell(\boldsymbol{\Psi}_{01}) = \sqrt{n}(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01})\{n^{-1}\nabla^2\ell(\boldsymbol{\Psi}_{01})\}$, we can therefore combine equations (S1.6) and (S1.7) to obtain

$$\frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\Psi}_{01}) = \sqrt{n}(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01})\mathcal{I}(\boldsymbol{\Psi}_{01}) + o_p(1)$$

The remainder of the proof follows a similar outline to the proof of Theorem 2 in Fan and Peng (2004). In particular, let

$$\boldsymbol{Y}_i = \frac{1}{\sqrt{n}}\boldsymbol{B}_n\mathcal{I}^{-1/2}(\boldsymbol{\Psi}_{01})\nabla\ell(\boldsymbol{\Psi}_{01}).$$

Then we can prove for $i = 1, \ldots, n$ that $\boldsymbol{Y}_i$ satisfies the Lindeberg condition. Application of the multivariate Lindeberg-Feller central limit theorem (Van der Vaart, 2000) leads to the results. □

S2 Additional Simulation Results

S2.1 Normal Responses

Table 1: Additional simulation results for linear mixed models. Performed was assessed in terms of the percentage of datasets where the correct model, i.e. both fixed and random effects structure, was chosen (%C), and the median relative model error (RME). Values of RME less than one indicates that CREPE has better model accuracy.

n	m	CREPE	M-ALASSO		ALASSO	
		%C	%C	RME	%C	RME
	5	23	17	0.66	0	1.01
30	10	74	67	0.96	15	0.52
	20	85	75	0.54	8	0.08
	5	50	29	0.67	11	1.10
60	10	89	69	0.90	41	0.99
	20	94	85	0.33	24	0.11

S2.2 Bernoulli Responses

Table 2: Additional simulation results for Bernoulli GLMMs. Performance was assessed in terms of the percentage of datasets where the correct model, i.e. both fixed and random effects structure, was chosen (%C), and the median relative model error (RME). Values of RME less than one indicate that CREPE has better model accuracy. Note %C for $\texttt{glmmLasso}_{\text{sat}}$ is zero by definition, and so its column is omitted from the model.

n	m	CREPE	$\texttt{glmmLasso}_{\text{true}}$		$\texttt{glmmLasso}_{\text{sat}}$
		%C	%C	RME	RME
50	10	5	29	2.09	1.91
	20	30	34	0.47	0.56
100	10	11	70	0.66	0.67
	20	51	76	0.58	0.58

S2.3 Poisson GLMMs

Datasets were simulated from a Poisson GLMM, using the same rate of growth of p as in Section 5.2 of the main text, i.e. $p = \lceil 7n^{1/4} \rceil$ where $\lceil \cdot \rceil$ is the ceiling function. Covariates $\boldsymbol{x}_{ij}$ were constructed with the first element

set to one for an intercept, and the remaining elements generated from a multivariate normal distribution with mean zero and covariance given by $\text{Cov}(x_{ijr}, x_{ijs}) = \rho^{|r-s|}$ and $\rho = 0.5$. The covariates for the random effects $\boldsymbol{z}_{ij}$ were taken as the first eight covariates of $\boldsymbol{x}_{ij}$. The first eight elements of $\boldsymbol{\beta}_0$ were set to $(0.5, 1, -1, 0, 1, 0, 0, 1)$, with the first element denoting the fixed intercept. Afterwards, every third element in $\boldsymbol{\beta}_0$ took alternating values of ± 1, while the remaining elements were set to zero. The true 8×8 covariance matrix $\boldsymbol{D}_0$ was structured as follows: 1) a 2×2 submatrix with diagonal elements 1 and off-diagonal elements of -0.5 occupied the top left of $\boldsymbol{D}_0$, 2) $[\boldsymbol{D}_0]_{88} = 1$, 3) all other elements were set to zero. Based on the above set up, responses y_{ij} were then generated from a Poisson distribution with log link. The response matrices generated had an average of 37% zero elements. We considered combinations of $n = 50, 100$ clusters, corresponding to $p = 19$ and 23 respectively, and cluster sizes of $m = 5, 10, 20$.

We compared CREPE (with $\nu = 2$ in the adaptive weights) with `glmmLasso` assuming either the random effects component was known and only elements 1, 2 and 8 of $\boldsymbol{z}_{ij}$ were included, or that it was unknown and the saturated random effects model was used. Furthermore, as was the case with the Bernoulli GLMM design, because `glmmLasso` only performs selection of the fixed effects, the model error was defined only in terms of

the fixed effects ME $= \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$.

CREPE performed strongly compared to the two versions of `glmmLasso` (Table 3): aside from the smallest sample size case of $(n, m) = (50, 5)$, it selected the correct random effects structure over 75% of the time. Furthermore, the mean number of false positives for the fixed effects dropped considerably when the cluster size increased from $m = 5$ to 20, while the mean number of false negatives was close to zero regardless of n and m. Finally, in all settings the median relative Kullback-Leibler distance for both versions of `glmmLasso` was smaller than one, indicating that CREPE had substantially better model accuracy (predictive capacity).

Table 3: Simulation results for Poisson GLMMs. Performance was assessed in terms of the mean number false positives (FP) and false negatives (FN) for the fixed effects, the percentage of datasets with correctly chosen random effects components (%RE, for CREPE only), the percentage of datasets where there was non-hierarchical shrinkage (%S), and median relative Kullback-Leibler distance (RKL). Values of RKL less than one indicated CREPE had better model accuracy. Since %S was equal to zero for all cases for CREPE, this column is omitted from the table.

n	m	CREPE			`glmmLasso`$_{\text{true}}$				`glmmLasso`$_{\text{sat}}$			
		FP	FN	%RE	FP	FN	%S	RKL	FP	FN	%S	RKL
50	5	2.44	0.11	52	3.12	2.99	52	0.25	2.78	3.44	80	0.40
	10	1.34	0.10	86	3.93	1.10	41	0.45	2.01	2.49	79	0.44
	20	0.40	0.09	89	3.73	0.49	19	0.83	1.10	2.06	54	0.61
100	5	2.34	0.08	77	2.78	3.69	71	0.13	3.00	3.33	77	0.39
	10	1.30	0.10	91	2.93	1.65	51	0.56	2.92	2.16	79	0.61
	20	0.36	0.08	92	4.89	0.65	19	0.76	3.62	1.45	43	0.72

The mean number of false positives was similar for both versions of `glmmLasso`, and both were substantially higher compared to CREPE. Also,

Table 4: Additional simulation results for Poisson GLMMs. Performance was assessed in terms of the percentage of datasets where the correct model, i.e. both fixed and random effects structure, was chosen (%C), and median relative model error (RME). Values of RME less than one indicate that CREPE has better model accuracy. Note %C for $\texttt{glmmLasso}_{\text{sat}}$ is zero by definition, and so its column is omitted from the model.

n	m	CREPE	$\texttt{glmmLasso}_{\text{true}}$		$\texttt{glmmLasso}_{\text{sat}}$
		%C	%C	RME	RME
50	5	7	0	0.627	0.44
	10	29	7	1.572	0.69
	20	65	31	2.306	1.05
100	5	9	1	0.406	0.45
	10	35	22	1.063	0.75
	20	72	27	1.044	0.92

the mean number of false negatives dropped dramatically with increasing cluster size for `glmmLasso`, although even at $m = 20$ it was still considerably higher than the CREPE estimator. Not surprisingly, assuming the true random effects structure led to a considerably smaller number of datasets with non-hierarchical shrinkage compared to assuming a saturated random effects structure. It was also not surprising to see that both the relative Kullback-Leibler distances and relative model errors were much closer to one for $\texttt{glmmLasso}_{\text{true}}$ than $\texttt{glmmLasso}_{\text{sat}}$, which reflects the fact that selecting only the truly non-zero random effects, rather than including all of them in the model, has implications for producing better estimates of the fixed effects. Interestingly however, at $m = 10$ and 20 the median RME for $\texttt{glmmLasso}_{\text{true}}$ was substantially greater than one, suggesting that

$\texttt{glmmLasso}_{\text{true}}$ outperformed CREPE in terms of estimating the fixed effect coefficients in these cases. While this result is confounded with the fact that CREPE performs joint selection and `glmmLasso` performed fixed effects selection only, subsequent investigation also showed that the estimated non-zero values of $\boldsymbol{\beta}$ from CREPE tended to be overshrunk, i.e. further from ± 1, compared to the $\texttt{glmmLasso}_{\text{true}}$ estimates, a result that needs further inquiry.

S3 R Code

The files `crepe-code.R` and `code-testing.R` contain `R` code for calculating the CREPE estimates for GLMMs and an example of how to use it respectively.

Bibliography

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.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics* **22**, 341–355.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.