Supplementary material for Maximum softly-penalized likelihood for Bernoulli-response generalized linear mixed models

Philipp Sterzinger[1] and Ioannis Kosmidis[1,2]

[1]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[2]The Alan Turing Institute, London, NW1 2DB, UK

March 11, 2022

S1 Supplementary Material

All labels for the sections, equations, tables, figures and so on in the current document have been prefixed by "S" (e.g. Section S1, equation (S21), etc). The supplementary material for *Maximum softly-penalized likelihood for Bernoulli-response generalized linear mixed models* contains:

i) Proofs of Theorems 4.1-4.3 of the main text and a generalization to approximate likelihoods (Section S2),

ii) Bounds of the first and second order partial derivatives of Jeffreys prior (Section S3),

iii) The multivariate extension of the negative Huber loss penalty and a proof that our proposed penalties give interior point estimates for the Bernoulli-response GLMM (Section S4),

iv) **#PS: Ioannis content**

v) A summary of the simulation study of Section 6 of the main paper, and

vi) Further simulations on synthetic data (Section S5)

This document and the R scripts and datasets to reproduce our results are available at `https://github.com/psterzinger/softpen_supplementary`. All estimations have been conducted in R version 3.6.3 (R Core Team, 2020) using the R-packages:

- `blme` (Chung et al., 2013) version 1.0-5
- `lme4` (Bates et al., 2015) version 1.1.27.1
- `numDeriv` (Gilbert and Varadhan, 2019) version 2016.8.1.1
- `optimx` (Nash and Varadhan, 2011) version 2021.6.12

S2 Asymptotic properties of the MSP(A)L

We recall our regularity assumptions for consistency.

SA0 Both $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ are differentiable, with derivatives $S(\boldsymbol{\theta}), \tilde{S}(\boldsymbol{\theta})$

SA1 $\sup_{\boldsymbol{\theta}\in\Theta} \left\| r_n^{-1} S(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta}) \right\| \xrightarrow{p} 0$ for some deterministic function $S_0(\boldsymbol{\theta})$

SA2 For all $\varepsilon > 0$, $\inf_{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\varepsilon} \|S_0(\boldsymbol{\theta})\| > 0 = \|S_0(\boldsymbol{\theta}_0)\|$

SA3 $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are roots of $S(\boldsymbol{\theta}), \tilde{S}(\tilde{\boldsymbol{\theta}})$, i.e. $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\tilde{S}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$.

Theorem S2.1 (Consistency)**:** *Let $\delta^\infty = o_p(r_n)$, and assume that SA0-SA3 hold. Then $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.*

Proof. The proof is analogous to the proof of Ogden (2017, Theorem 1) and follows Vaart (1998, Theorem 5.9). We give it here for completeness. First, let us bound $r_n^{-1} S(\tilde{\boldsymbol{\theta}})$:

$$\begin{aligned} \left\| r_n^{-1} S(\tilde{\boldsymbol{\theta}}) \right\| &= \left\| r_n^{-1} \tilde{S}(\tilde{\boldsymbol{\theta}}) + r_n^{-1} (S(\tilde{\boldsymbol{\theta}}) - \tilde{S}(\tilde{\boldsymbol{\theta}})) \right\| \\ &= \left\| \mathbf{0} - r_n^{-1} \nabla P(\tilde{\boldsymbol{\theta}}) \right\| \\ &= o_p(1) \end{aligned} \tag{S1}$$

where the second equality follows from the definition of $\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta})$ and SA3. The last equality follows from the assumption that $\delta^\infty = o_p(r_n)$. Hence, $\tilde{\boldsymbol{\theta}}$, in connection with SA0-SA2, satisfies the conditions of Theorem 5.9 in Vaart (1998), which guarantees consistency. In particular, by (S1), $\|r_n^{-1} S(\tilde{\boldsymbol{\theta}})\| = o_p(1)$ so that adding $\|S_0(\tilde{\boldsymbol{\theta}})\|$ to both sides of this equation and rearranging yields

$$\begin{aligned} \left\| S_0(\tilde{\boldsymbol{\theta}}) \right\| &= \left\| S_0(\tilde{\boldsymbol{\theta}}) \right\| - \left\| r_n^{-1} S(\tilde{\boldsymbol{\theta}}) \right\| + o_p(1) \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta} \left\| r_n^{-1} S(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta}) \right\| + o_p(1) \\ &= o_p(1) \end{aligned} \tag{S2}$$

where the second line follows from the reverse triangle inequality and the third from (S1) and SA1. Finally note that SA2 implies that for any $\varepsilon > 0$, there is a number η such that $\|S_0(\boldsymbol{\theta})\| > \eta$ for any $\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \varepsilon$. Hence, for any $\varepsilon > 0$, the event $\left\| \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\| \geq \varepsilon$ is implied by the event $\left\| S_0(\tilde{\boldsymbol{\theta}}) \right\| > \eta$, however, this was seen to converge to zero in probability in (S2). □

The proof of asymptotic normality follows the proof of Ogden (2017, Theorem 2), with the notation adapted to the soft penalization framework. Let us restate our assumptions for asymptotic normality.

SA4 Both $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ are three times differentiable

SA5 $\sup_{\boldsymbol{\theta}\in\Theta} \left\| \left| r_n^{-1} J(\boldsymbol{\theta}) - I(\boldsymbol{\theta}) \right| \right\| \xrightarrow{p} 0$ for some positive definite, nonrandom, $\mathcal{O}(1)$ matrix $I(\boldsymbol{\theta})$, that is continuous in $\boldsymbol{\theta}$ in a neighbourhood around $\boldsymbol{\theta}_0$

SA6 $r_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, I(\boldsymbol{\theta}_0)^{-1})$

SA7 $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$

The following Lemma, which is an adaptation of Lemma 1 in Ogden (2017), serves to relax the stochastic order requirement of $\nabla P(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ to achieve asymptotic normality.

Lemma S2.2. *Assume that SA3, SA4, SA5 and SA7 hold. Further suppose that* $\delta^\infty = o_p(r_n)$ *and that there is a* $t > 0$ *such that* $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(a_n)$ *for some nonnegative sequence* a_n *indexed by* n*. Then* $\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = o_p(r_n^{-1}a_n)$*.*

Proof. The proof is similar to Ogden (2017, Lemma 1). A first order Taylor expansion of $S(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ yields

$$r_n^{-1}S(\boldsymbol{\theta}) = r_n^{-1}S(\hat{\boldsymbol{\theta}}) + r_n^{-1}\nabla S(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = -J(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{S3}$$

where $\boldsymbol{\theta}^*$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

Now since $\nabla P(\boldsymbol{\theta}) = \tilde{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})$, substituting $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in (S3), gives that

$$\mathbf{0} = r_n^{-1}\tilde{S}(\tilde{\boldsymbol{\theta}}) = r_n^{-1}S(\tilde{\boldsymbol{\theta}}) + r_n^{-1}\nabla P(\tilde{\boldsymbol{\theta}}) = -r_n^{-1}J(\boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + r_n^{-1}\nabla P(\tilde{\boldsymbol{\theta}}) \tag{S4}$$

so that

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = [r_n^{-1}J(\boldsymbol{\theta}^*)]^{-1}r_n^{-1}\nabla P(\tilde{\boldsymbol{\theta}}) \tag{S5}$$

for some $\boldsymbol{\theta}^*$ between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. Now by SA7, $\tilde{\boldsymbol{\theta}}$ is consistent so that also $\boldsymbol{\theta}^*$ is consistent for $\boldsymbol{\theta}_0$. Hence by assumption SA5, it follows that $[r_n^{-1}J(\boldsymbol{\theta}^*)]^{-1}$ converges in probability to $I(\boldsymbol{\theta}_0)^{-1}$ so that $\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = \mathcal{O}_p(r_n^{-1}\delta(\tilde{\boldsymbol{\theta}}))$.

Now let $A_t = \{\tilde{\boldsymbol{\theta}} \in B_t(\boldsymbol{\theta}_0)\}$ for the t such that $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(a_n)$ for some nonnegative sequence a_n indexed by n. Denote by $\bar{A}_t$, the complement to A_t. Then, by construction, conditional on A_t,

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = \mathcal{O}_p(r_n^{-1}\delta^\infty(B_t(\boldsymbol{\theta}_0))) = o_p(r_n^{-1}a_n)\,.$$

Moreover, since $\tilde{\boldsymbol{\theta}}$ is consistent, $\Pr(A_t) \to 1$ as $n \to \infty$. Putting everything together, one gets that for any $\varepsilon > 0$,

$$\begin{aligned}\Pr\left(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \geq \varepsilon r_n^{-1}a_n\right) &= \Pr\left(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \geq \varepsilon r_n^{-1}a_n \mid A_t\right)\Pr(A_t) \\ &\quad + \Pr\left(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \geq \varepsilon r_n^{-1}a_n \mid \bar{A}_t\right)\Pr(\bar{A}_t) \\ &\leq \Pr\left(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \geq \varepsilon r_n^{-1}a_n \mid A_t\right) + \Pr(\bar{A}_t) \to 0, \quad n \to \infty\end{aligned} \tag{S6}$$

as required. □

An immediate consequence of the Lemma and assumptions SA3-SA7 is the following theorem.

Theorem S2.3 (Asymptotic Normality)**:** *Assume that conditions SA3-SA7 hold. Let* $\delta^\infty = o_p(r_n)$ *and assume there is a* $t > 0$ *such that* $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$*. Then* $r_n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, I(\boldsymbol{\theta}_0)^{-1})$*.*

Proof. By Lemma S2.2, it holds that $\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = o_p(r_n^{-1/2})$ so that $r_n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = r_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + r_n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = r_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$ and the result follows by SA6. □

Theorem S2.4 (Hypothesis testing)**:** *Assume that conditions SA3-SA7 hold and that* $\delta^\infty = o_p(r_n)$*,* $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$ *and* $\gamma^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n)$ *for some* $t > 0$*. Then, under* $H_0 : \tilde{\Lambda} - \Lambda = o_p(1)$*.*

Proof. We follow the proof of Ogden (2017, Theorem 3). By definition of $\Lambda, \tilde{\Lambda}$, we have

$$\frac{\tilde{\Lambda} - \Lambda}{2} = \{\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\tilde{\boldsymbol{\theta}}^R)\} - \{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}^R)\} \tag{S7}$$

Adding and subtracting $\tilde{\ell}(\hat{\boldsymbol{\theta}})$ and $\tilde{\ell}(\hat{\boldsymbol{\theta}}^R)$ and rearranging yields

$$\frac{\tilde{\Lambda} - \Lambda}{2} = \{\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\hat{\boldsymbol{\theta}})\} + \{\tilde{\ell}(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}})\} + \{\tilde{\ell}(\hat{\boldsymbol{\theta}}^R) - \tilde{\ell}(\tilde{\boldsymbol{\theta}}^R)\} + \{\ell(\hat{\boldsymbol{\theta}})^R - \tilde{\ell}(\hat{\boldsymbol{\theta}}^R)\} \tag{S8}$$

We first bound $\{\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\hat{\boldsymbol{\theta}})\}$. For this, we do a second order Taylor expansion of $\tilde{\ell}(\hat{\boldsymbol{\theta}})$ around $\tilde{\boldsymbol{\theta}}$. Upon rearranging, we get that

$$\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\hat{\boldsymbol{\theta}}) = -\tilde{S}(\tilde{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top \tilde{J}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = \frac{1}{2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top \tilde{J}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \tag{S9}$$

where $\boldsymbol{\theta}^*$ lies on the line segment between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, i.e. $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}} + c(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$, $c \in [0,1]$ and $\tilde{J}(\boldsymbol{\theta}) = -\nabla\nabla^\top \tilde{\ell}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) - \nabla\nabla^\top P(\boldsymbol{\theta})$. We next show that $\tilde{J}(\boldsymbol{\theta}^*) = \mathcal{O}_p(r_n)$, by establishing that $\nabla\nabla^\top P(\boldsymbol{\theta}^*) = o_p(r_n)$ and that $J(\boldsymbol{\theta}^*) = \mathcal{O}_p(r_n)$. For this, let A_t be the event that $\boldsymbol{\theta}^* \in B_t(\boldsymbol{\theta}_0)$ and denote by $\bar{A}_t$ its complement. Then for any $\varepsilon > 0$,

$$\begin{aligned}
\Pr\left(|||r_n^{-1}\nabla\nabla^\top P(\boldsymbol{\theta}^*)||| < \varepsilon\right) &= \Pr\left(|||r_n^{-1}\nabla\nabla^\top P(\boldsymbol{\theta}^*)||| < \varepsilon \mid A_t\right)\Pr(A_t) \\
&\quad + \Pr\left(|||r_n^{-1}\nabla\nabla^\top P(\boldsymbol{\theta}^*)||| < \varepsilon \mid \bar{A}_t\right)\Pr(\bar{A}_t) \\
&\leq \Pr\left(|||r_n^{-1}\nabla\nabla^\top P(\boldsymbol{\theta}^*)||| < \varepsilon \mid A_t\right) + \Pr(\bar{A}_t) \to 0, \quad n \to \infty
\end{aligned} \tag{S10}$$

where the first equality follows from the definition of conditional probabilities, and the second line from the fact that all probabilities are between zero and one. By assumption SA7, it follows that $\Pr(\bar{A}_t)$ converges to zero and by the assumption that $\gamma^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n)$, it follows that $\Pr\left(|||r_n^{-1}\nabla\nabla^\top P(\boldsymbol{\theta}^*)||| < \varepsilon \mid A_t\right) \leq \Pr(r_n^{-1}\gamma^\infty(B_t(\boldsymbol{\theta}_0)) > \varepsilon)$, which converges to zero. By a similar argument and using SA5, it holds that $J(\boldsymbol{\theta}^*) = \mathcal{O}_p(r_n)$ so that we conclude that indeed $\tilde{J}(\boldsymbol{\theta}^*) = \mathcal{O}_p(r_n)$. By Lemma S2.2, we know that $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = o_p(r_n^{-1/2})$ and thus

$$\frac{1}{2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top \tilde{J}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = o_p(1) \tag{S11}$$

Moreover, by SA3, $\tilde{S}(\tilde{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = \mathbf{0}$, and therefore $\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\hat{\boldsymbol{\theta}}) = o_p(1)$. A similar argument shows that under H_0, $\tilde{\ell}(\hat{\boldsymbol{\theta}}^R) - \tilde{\ell}(\tilde{\boldsymbol{\theta}}^R) = o_p(1)$. Hence, (S8) becomes

$$\frac{\tilde{\Lambda} - \Lambda}{2} = P(\hat{\boldsymbol{\theta}}) - P(\hat{\boldsymbol{\theta}}^R) + o_p(1) \tag{S12}$$

Now a first order Taylor expansion of $P(\hat{\boldsymbol{\theta}})$ around $\hat{\boldsymbol{\theta}}^R$ yields

$$\frac{\tilde{\Lambda} - \Lambda}{2} = \nabla P(\boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^R) + o_p(1) \tag{S13}$$

for some $\boldsymbol{\theta}^*$ on the line segment between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^R$. Now under H_0, $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^R = \mathcal{O}_p(r_n^{-1/2})$ Again, let A_t be the event that $\boldsymbol{\theta}^* \in B_t(\boldsymbol{\theta}_0)$ and denote by $\bar{A}_t$ its complement. Then for any $\varepsilon > 0$

$$\begin{aligned}
\Pr(r_n^{-1/2}\|\nabla P(\boldsymbol{\theta}^*)\| > \varepsilon) &\leq \Pr(r_n^{-1/2}\|\nabla P(\boldsymbol{\theta}^*)\| > \varepsilon \mid A_t)\Pr(A_t) \\
&\quad + \Pr(r_n^{-1/2}\|\nabla P(\boldsymbol{\theta}^*)\| > \varepsilon \mid \bar{A}_t)\Pr(\bar{A}_t) \\
&\leq \Pr(r_n^{-1/2}\|\nabla P(\boldsymbol{\theta}^*)\| > \varepsilon \mid A_t) + \Pr(\bar{A}_t) \\
&\leq \Pr(r_n^{-1/2}\delta^\infty(B_t(\boldsymbol{\theta}_0) > \varepsilon) + \Pr(\bar{A}_t) \to 0, \quad n \to \infty
\end{aligned} \tag{S14}$$

so that $\nabla P(\boldsymbol{\theta}^*) = o_p(r_n^{1/2})$ and thus $\frac{\tilde{\Lambda}-\Lambda}{2} = o_p(1)$ as required. □

If $\ell(\boldsymbol{\theta})$ refers to an exact model likelihood that is unavailable, the framework of Ogden (2017) readily gives conditions on the approximation error that preserve the asymptotic properties of Theorems S2.1-S2.4 for $\tilde{\boldsymbol{\theta}}$. Let $\bar{\ell}(\boldsymbol{\theta})$ be an approximation to $\ell(\boldsymbol{\theta})$ and denote by $\bar{S}(\boldsymbol{\theta})$ its score and by $\bar{J}(\boldsymbol{\theta})$ its observed information matrix. For $S \subseteq \Theta$, let $\bar{\delta}^\infty(S) = \sup_{\boldsymbol{\theta}\in S} \|\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})\|$ and $\bar{\delta}^\infty = \bar{\delta}^\infty(\Theta)$. Let $\tilde{\ell}(\boldsymbol{\theta}) = \bar{\ell}(\boldsymbol{\theta}) + P(\boldsymbol{\theta})$ and denote by $\tilde{\boldsymbol{\theta}}$ its maximizer over Θ. Finally, define $\bar{\gamma}^\infty(S) = \sup_{\boldsymbol{\theta}\in S} |||J(\boldsymbol{\theta}) - \bar{J}(\boldsymbol{\theta})|||$.

Corollary S2.1 (Consistency of MSPAL estimates)**.** *Assume that assumptions SA0-SA3 hold. Further assume that* $\delta^\infty = o_p(r_n)$ *and that* $\bar{\delta}^\infty = o_p(r_n)$*. Then* $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$*.*

Proof. Note that since $\tilde{S}(\boldsymbol{\theta}) = S(\boldsymbol{\theta}) + (\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})) + \nabla P(\boldsymbol{\theta})$, by the assumptions of the Corollary and SA3, it follows that

$$\mathbf{0} = r_n^{-1}\tilde{S}(\tilde{\boldsymbol{\theta}}) = r_n^{-1}S(\tilde{\boldsymbol{\theta}}) + r_n^{-1}(\bar{S}(\tilde{\boldsymbol{\theta}}) - S(\tilde{\boldsymbol{\theta}})) + r_n^{-1}\nabla P(\tilde{\boldsymbol{\theta}}) \tag{S15}$$

and thus

$$r_n^{-1}S(\tilde{\boldsymbol{\theta}}) = -r_n^{-1}(\bar{S}(\tilde{\boldsymbol{\theta}}) - S(\tilde{\boldsymbol{\theta}})) - r_n^{-1}\nabla P(\tilde{\boldsymbol{\theta}}) = o_p(1) \tag{S16}$$

Now the argument of Theorem S2.1 can be applied from (S1) onwards. □

Corollary S2.2 (Asymptotic normality of MSPAL estimates)**.** *Assume that conditions SA3-SA7 hold. Let* $\delta^\infty = o_p(r_n), \bar{\delta}^\infty = o_p(r_n)$ *and assume there is a* $t > 0$ *such that* $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2}), \bar{\delta}^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$*. Then* $r_n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, I(\boldsymbol{\theta}_0)^{-1})$*.*

Proof. Similar to the proof of Corollary S2.1, define a new "penalty" $\bar{P}(\boldsymbol{\theta}) = \{\bar{\ell}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})\} + P(\boldsymbol{\theta})$. Then by the triangle inequality,

$$\begin{aligned} \|\nabla\bar{P}(\boldsymbol{\theta})\| &= \|\{\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})\} + \nabla P(\boldsymbol{\theta})\| \\ &\leq \|\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})\| + \|\nabla P(\boldsymbol{\theta})\| \end{aligned} \tag{S17}$$

and therefore,

$$\begin{aligned} \sup_{\boldsymbol{\theta}\in\Theta}\|\nabla\bar{P}(\boldsymbol{\theta})\| &\leq \sup_{\boldsymbol{\theta}\in\Theta}\|\{\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})\| &&+ \sup_{\boldsymbol{\theta}\in\Theta}\|\nabla P(\boldsymbol{\theta})\| &&= o_p(r_n) \\ \sup_{\boldsymbol{\theta}\in B_t(\boldsymbol{\theta}_0)}\|\nabla\bar{P}(\boldsymbol{\theta})\| &\leq \sup_{\boldsymbol{\theta}\in B_t(\boldsymbol{\theta}_0)}\|\{\bar{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta})\| &&+ \sup_{\boldsymbol{\theta}\in B_t(\boldsymbol{\theta}_0)}\|\nabla P(\boldsymbol{\theta})\| &&= o_p(r_n^{1/2}) \end{aligned} \tag{S18}$$

Hence, the proof of Theorem S2.3 applies for $\bar{P}(\boldsymbol{\theta})$ in place of $P(\boldsymbol{\theta})$. □

Corollary S2.3 (Hypothesis testing for of MSPAL)**.** *Assume that conditions SA3-SA7 hold and that* $\delta^\infty = o_p(r_n), \bar{\delta}^\infty = o_p(r_n)$*,* $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$*,* $\bar{\delta}^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$ *and* $\gamma^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n), \bar{\gamma}^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n)$ *for some* $t > 0$*. Then, under* $H_0 : \tilde{\Lambda} - \Lambda = o_p(1)$*.*

Proof. Similarly to the proof of Corollary S2.2, define $\bar{P}(\boldsymbol{\theta}) = \{\bar{\ell}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})\} + P(\boldsymbol{\theta})$. Then by virtue of the triangle inequality, and the assumptions of the Corollary,

$$\begin{aligned} \sup_{\sup\boldsymbol{\theta}\in\Theta}\|\nabla\bar{P}(\boldsymbol{\theta})\| &= o_p(r_n) \\ \sup_{\sup\boldsymbol{\theta}\in B_t(\boldsymbol{\theta}_0)}\|\nabla\bar{P}(\boldsymbol{\theta})\| &= o_p(r_n^{1/2}) \\ \sup_{\sup\boldsymbol{\theta}\in B_t(\boldsymbol{\theta}_0)}\|\nabla\nabla^\top\bar{P}(\boldsymbol{\theta})\| &= o_p(r_n) \end{aligned} \tag{S19}$$

Hence, the proof of Theorem S2.4 applies for $\bar{P}(\boldsymbol{\theta})$ in place of $P(\boldsymbol{\theta})$. □

S3 Bounds on the derivatives Jeffreys prior

In this section we give bounds on the first and second partial derivatives of the logarithm of Jeffreys invariant prior from a Bernoulli-response GLM with logistic link. We expect the derivations to be useful for similar such results under different link functions.

Theorem S3.1 (Bounding the partial derivative of the log of Jeffreys invariant prior)**:** *Let* $\boldsymbol{X} \in \Re^{n\times p}$ *be a full column rank matrix,* $\boldsymbol{W}$ *a diagonal matrix with entries* $w_j = [\boldsymbol{W}]_{jj} = \mu_j(\boldsymbol{\beta})(1-\mu_j(\boldsymbol{\beta}))$ *and* $logit(\mu_j(\boldsymbol{\beta})) = \boldsymbol{x}_j^\top \boldsymbol{\beta}$, $\boldsymbol{\beta} \in \Re^p$. *Then*

$$\left|\frac{\partial}{\partial \beta_i} \log\det(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})\right| \leq p \max_{1\leq j\leq n} |x_{ji}(1-2\mu_j(\boldsymbol{\beta}))| \leq p \max_{1\leq j\leq n} |x_{ji}|, \quad \textit{and} \tag{S20}$$

$$\left|\frac{\partial^2}{\partial \beta_i \partial \beta_j} \log\det(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})\right| \leq 2p \max_{1\leq k\leq n} |x_{ki}| \max_{1\leq k\leq n} |x_{kj}| \tag{S21}$$

Proof. We first consider the first partial derivative. It is noted without proof that

$$\left|\frac{\partial}{\partial \beta_i} \log\det(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})\right| = \operatorname{tr}\left((\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\widetilde{\boldsymbol{W}}_i\boldsymbol{X}\right) \tag{S22}$$

where $\widetilde{\boldsymbol{W}}_i$ is a diagonal matrix with diagonal entries $\widetilde{w}_j^{(i)} = [\widetilde{\boldsymbol{W}}_i]_{jj} = x_{ji}(1-2\mu_j(\boldsymbol{\beta}))$. Now by the cyclical property of the trace, it follows that

$$\operatorname{tr}\left((\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\widetilde{\boldsymbol{W}}_i\boldsymbol{X}\right) = \operatorname{tr}\left(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\widetilde{\boldsymbol{W}}_i\right) \tag{S23}$$

For notational brevity, denote the projection matrix $\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}$ by $\boldsymbol{P}$. Since $\widetilde{\boldsymbol{W}}_i$ is a diagonal matrix, one gets that

$$\begin{aligned}
\left|\operatorname{tr}\left(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\widetilde{\boldsymbol{W}}_i\right)\right| &= \left|\sum_{j=1}^n \widetilde{w}_j^{(i)}[\boldsymbol{P}]_{jj}\right| && \text{(S24)}\\
&\leq \sum_{j=1}^n \left|\widetilde{w}_j^{(i)}[\boldsymbol{P}]_{jj}\right| && \text{(S25)}\\
&\leq \sum_{j=1}^n \left|\widetilde{w}_j^{(i)}\right| [\boldsymbol{P}]_{jj} && \text{(S26)}\\
&\leq \max_{1\leq j\leq n} \left|\widetilde{w}_j^{(i)}\right| \sum_{j=1}^n [\boldsymbol{P}]_{jj} && \text{(S27)}\\
&= p \max_{1\leq j\leq n} \left|\widetilde{w}_j^{(i)}\right| && \text{(S28)}\\
&= p \max_{1\leq j\leq n} |x_{ji}(1-2\mu_j(\boldsymbol{\beta}))| && \text{(S29)}\\
&\leq p \max_{1\leq j\leq n} |x_{ji}| && \text{(S30)}
\end{aligned}$$

Here the second line is due to the triangle inequality. The third line follows by positive-semi-definiteness of $\boldsymbol{P}$ and the well known property that the main-diagonal entries of a positive-semi-definite matrix are nonnegative. To see that $\boldsymbol{P}$ is positive-semi definite, note that $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$ is positive definite as $\boldsymbol{X}$ has full column rank and $\boldsymbol{W}$ is a diagonal matrix with positive entries. It thus follows that $(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}$ is positive definite. Hence for any $\boldsymbol{y} \in \Re^n, \|\boldsymbol{y}\|_2 \neq 0$, $\boldsymbol{y}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \tilde{\boldsymbol{y}}^\top (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\tilde{\boldsymbol{y}} \geq 0$, for $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{y}$. Hence, as $\boldsymbol{W}$ is a diagonal matrix with nonnegative diagonal entries it follows that the main diagonal entries of $\boldsymbol{P}$, which are the elementwise product of the diagonals of $\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}$ and $\boldsymbol{W}$ are nonnegative. The fifth line follows since $\boldsymbol{P}$ is an idempotent matrix of rank p, and the fact that the trace of an idempotent matrix equals its rank (Harville, 1998, Corollary 10.2.2). The fact that $\boldsymbol{P}$ has rank p follows from the assumption that $\boldsymbol{X}$ has full column rank and since $\boldsymbol{W}$ is invertible for any $\boldsymbol{\beta} \in \Re^p, \boldsymbol{X} \in \Re^{n\times p}$ by construction and is a standard result in linear algebra (see for example Magnus and Neudecker (2019), Chapter 1.7). The last line follows since $\mu_j(\boldsymbol{\beta}) = \text{logit}^{-1}(\boldsymbol{x}_j^\top \boldsymbol{\beta}) \in (0,1)$.

Now consider the second partial derivative of $\log\det(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})$ with respect to β_i, β_j. That is,

$$\begin{aligned}\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}) &= \frac{\partial}{\partial\beta_j}\mathrm{tr}\left(\underbrace{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{:=\boldsymbol{S}}\boldsymbol{W}\widetilde{\boldsymbol{W}}_i\right)\\ &= \frac{\partial}{\partial\beta_j}\sum_{k=1}^n \boldsymbol{S}_{kk}w_k\widetilde{w}_k^{(i)} \\ &= \sum_{k=1}^n\left[\frac{\partial}{\partial\beta_j}\boldsymbol{S}_{kk}\right]w_k\widetilde{w}_k^{(i)} + \sum_{k=1}^n \boldsymbol{S}_{kk}\left[\frac{\partial}{\partial\beta_j}w_k\widetilde{w}_k^{(i)}\right]\end{aligned} \tag{S31}$$

First consider $\frac{\partial}{\partial\beta_j}\boldsymbol{S}_{kk}$:

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\boldsymbol{S} &= \frac{\partial}{\partial\beta_j}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \\ &= \boldsymbol{X}\left[\frac{\partial}{\partial\beta_j}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\right]\boldsymbol{X}^\top \\ &= \boldsymbol{X}\left[-(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\widetilde{\boldsymbol{W}}_j\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\right]\boldsymbol{X}^\top\end{aligned} \tag{S32}$$

Now letting $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}$, one gets that

$$\begin{aligned}\sum_{k=1}^n\left[\frac{\partial}{\partial\beta_j}\boldsymbol{S}_{kk}\right]w_k\widetilde{w}_k^{(i)} &= -\mathrm{tr}\left(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\widetilde{\boldsymbol{W}}_j\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\widetilde{\boldsymbol{W}}_i\right)\\ &= -\mathrm{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right)\end{aligned} \tag{S33}$$

Now recall that $\mathrm{logit}(\mu_k) = \boldsymbol{x}_k^\top\boldsymbol{\beta}$ and consider $\frac{\partial}{\partial\beta_j}w_k\widetilde{w}_k^{(i)}$:

$$\begin{aligned}\frac{\partial}{\partial\beta_j}w_k\widetilde{w}_k^{(i)} &= \frac{\partial}{\partial\beta_j}\mu_k(1-\mu_k)(1-2\mu_k)x_{ki} \\ &= \mu_k(1-\mu_k)\left[(1-2\mu_k)(1-3\mu_k)-\mu_k\right]x_{ki}x_{kj}\end{aligned} \tag{S34}$$

Letting $\widetilde{\widetilde{\boldsymbol{W}}}_{ij}$ be a diagonal matrix with diagonal entries $[\widetilde{\widetilde{\boldsymbol{W}}}_{ij}]_{kk} = [(1-2\mu_k)(1-3\mu_k)-\mu_k]\,x_{ki}x_{kj}$, it thus follows that

$$\sum_{k=1}^n \boldsymbol{S}_{kk}\left[\frac{\partial}{\partial\beta_j}w_k\widetilde{w}_k^{(i)}\right] = \mathrm{tr}\left(\boldsymbol{P}\widetilde{\widetilde{\boldsymbol{W}}}_{ij}\right) \tag{S35}$$

and therefore

$$\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}) = \mathrm{tr}\left(\boldsymbol{P}\widetilde{\widetilde{\boldsymbol{W}}}_{ij}\right) - \mathrm{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right) \tag{S36}$$

Towards bounding these expressions, note that by similar arguments to (S24)-(S30), it holds that

$$\left|\mathrm{tr}\left(\boldsymbol{P}\widetilde{\widetilde{\boldsymbol{W}}}_{ij}\right)\right| \le p\max_{1\le k\le n}\left|\left[(1-2\mu_k)(1-3\mu_k)-\mu_k\right]x_{ki}x_{kj}\right| \le p\max_{1\le k\le n}|x_{ki}x_{kj}| \tag{S37}$$

where the last inequality follows since $(1-2\mu_k)(1-3\mu_k)-\mu_k \in [-\frac{1}{2}, 1]$.

Now towards bounding $\mathrm{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right)$, we make the following observations.

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right) &= \sum_{k=1}^{n}\left[\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right]_{kk} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n}\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\widetilde{w}_l^{(j)}\widetilde{w}_k^{(i)}\end{aligned} \tag{S38}$$

Also note that

$$\begin{aligned}\sum_{l=1}^{n}\left|\widetilde{w}_k^{(j)}\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\right| &= \sum_{l=1}^{n}\left|\widetilde{w}^{(j)}w_kw_l[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top]_{lk}^2\right| \\ &= \sum_{l=1}^{n}\left|\widetilde{w}^{(j)}\right|w_kw_l[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top]_{lk}^2 \\ &= \sum_{l=1}^{n}|\widetilde{w}^{(j)}|\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\end{aligned} \tag{S39}$$

Hence

$$\begin{aligned}\left|\operatorname{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right)\right| &= \left|\sum_{k=1}^{n}\left[\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right]_{kk}\right| \\ &= \left|\sum_{k=1}^{n}\widetilde{w}_k^{(i)}\sum_{l=1}^{n}\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\widetilde{w}_l^{(j)}\right| \\ &\le \sum_{k=1}^{n}\left|\widetilde{w}_k^{(i)}\right|\sum_{l=1}^{n}\left|\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\widetilde{w}_l^{(j)}\right| \\ &\le \max_{1\le k\le n}\left|\widetilde{w}_k^{(i)}\right|\sum_{k=1}^{n}\sum_{l=1}^{n}\boldsymbol{P}_{kl}\boldsymbol{P}_{lk}\left|\widetilde{w}_l^{(j)}\right| \\ &= \max_{1\le k\le n}\left|\widetilde{w}_k^{(i)}\right|\operatorname{tr}\left(\boldsymbol{P}|\widetilde{\boldsymbol{W}}_j|\boldsymbol{P}\right) \\ &= \max_{1\le k\le n}\left|\widetilde{w}_k^{(i)}\right|\operatorname{tr}\left(\boldsymbol{P}|\widetilde{\boldsymbol{W}}_j|\right) \\ &\le p\max_{1\le k\le n}\left|\widetilde{w}_k^{(i)}\right|\max_{1\le k\le n}\left|\widetilde{w}_k^{(j)}\right| \\ &= p\max_{1\le k\le n}|x_{ki}(1-2\mu_k)|\max_{1\le k\le n}|x_{kj}(1-2\mu_k)| \\ &\le p\max_{1\le k\le n}|x_{ki}|\max_{1\le k\le n}|x_{kj}|\end{aligned} \tag{S40}$$

So that it follows that

$$\begin{aligned}\left|\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})\right| &= \left|\operatorname{tr}\left(\boldsymbol{P}\widetilde{\widetilde{\boldsymbol{W}}}_{ij}\right)-\operatorname{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right)\right| \\ &\le \left|\operatorname{tr}\left(\boldsymbol{P}\widetilde{\widetilde{\boldsymbol{W}}}_{ij}\right)\right|+\left|\operatorname{tr}\left(\boldsymbol{P}\widetilde{\boldsymbol{W}}_j\boldsymbol{P}\widetilde{\boldsymbol{W}}_i\right)\right| \\ &\le p\max_{1\le k\le n}|x_{ki}x_{kj}|+p\max_{1\le k\le n}|x_{ki}|\max_{1\le k\le n}|x_{kj}| \\ &\le 2p\max_{1\le k\le n}|x_{ki}|\max_{1\le k\le n}|x_{kj}|\end{aligned} \tag{S41}$$

This concludes the proof. □

S4 Interior point parameter estimates

We first show that any lower triangular matrix $\tilde{\boldsymbol{L}}$ with finite entries and strictly positive entries on its main diagonal defines a nondegenerate variance covariance matrix $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$. Using this result, we show that our proposed penalty gives a nondegenerate variance components estimate and finite fixed effects estimates for a Bernoulli-response GLMM.

S4.1 Nondegenerate variance components estimates through Huber loss penalty

Let $\boldsymbol{\Sigma} \in \Re^{q\times q}$ be a real, symmetric, positive definite (variance-covariance) matrix and denote its unique lower triangular Cholesky factor by $\boldsymbol{L}$. Suppose we estimate $\boldsymbol{L}$, denote its estimate by $\tilde{\boldsymbol{L}}$, to obtain an estimate $\tilde{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$, based on $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$. It is assumed that $\tilde{\boldsymbol{L}}$ is a real, lower triangular matrix. We say that $\tilde{\boldsymbol{\Sigma}}$ is degenerate if one of the following holds:

S1) $\tilde{\boldsymbol{\Sigma}}$ is not symmetric

S2) $\tilde{\boldsymbol{\Sigma}}$ is not positive definite

S3) Some entries of $\tilde{\boldsymbol{\Sigma}}$ are infinite in absolute value

S4) There is perfect estimated correlation, i.e. there exist indices $i, j : i \neq j$: $\left|\frac{\tilde{\boldsymbol{\Sigma}}_{ij}}{\sqrt{\tilde{\boldsymbol{\Sigma}}_{ii}\tilde{\boldsymbol{\Sigma}}_{jj}}}\right| = 1$

Lemma S4.1. *Let $\tilde{\boldsymbol{L}} \in \Re^{q\times q}$ be real, lower triangular matrix with finite entries and strictly positive entries on its main diagonal. Then $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$ is not degenerate.*

Proof. S1) is trivial and follows from $\tilde{\boldsymbol{\Sigma}}^\top = (\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top)^\top = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top = \tilde{\boldsymbol{\Sigma}}$. To show that S2) cannot hold, take any $\boldsymbol{x} \in \Re^q$, $\boldsymbol{x} \neq \boldsymbol{0}_q$. Then by straightforward manipulations

$$\begin{aligned} \boldsymbol{x}^\top\tilde{\boldsymbol{\Sigma}}\boldsymbol{x} &= \boldsymbol{x}^\top\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top\boldsymbol{x} \\ &= (\tilde{\boldsymbol{L}}^\top\boldsymbol{x})^\top\tilde{\boldsymbol{L}}^\top\boldsymbol{x} \\ &= \langle \boldsymbol{y}, \boldsymbol{y}\rangle, \quad \boldsymbol{y} = \tilde{\boldsymbol{L}}^\top\boldsymbol{x} \\ &\geq 0 \end{aligned} \tag{S42}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean inner product. Hence $\tilde{\boldsymbol{\Sigma}}$ is positive semidefinite. Suppose that there is some $\boldsymbol{x} \in \Re^d$ such that $\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x} = 0$. Then by (S42), $\langle \boldsymbol{y}, \boldsymbol{y}\rangle = 0$ which holds if and only if $\boldsymbol{y} = \boldsymbol{0}_d$. Now since $\tilde{\boldsymbol{L}}$ is lower triangular with strictly positive diagonal entries, it is full rank. To see this, assume that $\tilde{\boldsymbol{L}}\boldsymbol{x} = \boldsymbol{0}_d$ and note that $[\tilde{\boldsymbol{L}}\boldsymbol{x}]_1 = l_{11}x_1$. Since $\tilde{l}_{11} > 0$, it must be that $x_1 = 0$. Now $[\tilde{\boldsymbol{L}}\boldsymbol{x}]_2 = \tilde{l}_{21}x_1 + \tilde{l}_{22}x_2$. Since $x_1 = 0$ and $\tilde{l}_{22} > 0$ it must again hold that $x_2 = 0$ and by induction $\boldsymbol{x} = \boldsymbol{0}_d$ so that $\tilde{\boldsymbol{L}}$ is full rank. But then $\boldsymbol{y} = \tilde{\boldsymbol{L}}\boldsymbol{x} = \boldsymbol{0}_d$ implies that $\boldsymbol{x} = \boldsymbol{0}_d$ so that $\tilde{\boldsymbol{\Sigma}}$ is positive definite. To prove that S3) does not hold, note that $\tilde{\boldsymbol{\Sigma}}_{ij} = \langle\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{l}}_j\rangle$, where $\tilde{\boldsymbol{l}}_i$ is the ith row vector of $\tilde{\boldsymbol{L}}$. Since all elements of $\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{l}}_j$ are finite, so is their inner product. Finally, towards a contradiction, assume that S4) holds. Then there exist some indices $i, j, i \neq j$ such that

$$\left|\frac{\tilde{\boldsymbol{\Sigma}}_{ij}}{\sqrt{\tilde{\boldsymbol{\Sigma}}_{ii}\tilde{\boldsymbol{\Sigma}}_{jj}}}\right| = 1 \tag{S43}$$

$$\iff \quad |\tilde{\boldsymbol{\Sigma}}_{ij}| = \sqrt{\tilde{\boldsymbol{\Sigma}}_{ii}\tilde{\boldsymbol{\Sigma}}_{jj}} \tag{S44}$$

$$\iff \quad |\langle\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{l}}_j\rangle| = \|\tilde{\boldsymbol{l}}_i\|\|\tilde{\boldsymbol{l}}_j\| \tag{S45}$$

where $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ is the induced inner product norm. It follows from the Cauchy-Schwarz inequality that the equality in the last line of (S43) holds if and only if $\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{l}}_j$ are linearly dependent. Since $\tilde{\boldsymbol{L}}$ is lower triangular, this is only possible if $\tilde{\boldsymbol{l}}_i, \tilde{\boldsymbol{l}}_j$ have zeroes in the same positions. But since all diagonal entries of $\tilde{\boldsymbol{L}}$ are strictly positive, this is not possible. Hence S4) cannot hold. □

S4.2 Interior point parameter estimates for Bernoulli-response GLMMS

Theorem S4.2 (Interior point estimates)**:** *Let $\ell(\boldsymbol{\theta})$ be the log-likelihood of a Bernoulli response GLMM, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{L})$ and $\boldsymbol{L}$ is the Cholesky factor of the variance components matrix $\boldsymbol{\Sigma}$. Let*

$$\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{L}}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \{\ell(\boldsymbol{\theta}) + P^{FE}(\boldsymbol{\beta}) + P^{MVRE}(\boldsymbol{L})\}, \tag{S46}$$

be the maximizer of the penalized model log-likelihood, with

$$P^{MVRE}(\boldsymbol{L}) = \sum_{i=1}^{N_q} P^{RE}(\log(l_{ii})) + \sum_{i<j}^{q} P^{RE}(l_{ij}), \tag{S47}$$

$$P^{RE}(x) \propto \begin{cases} -\frac{1}{2}\{x\}^2, & \text{if } |x| \leq 1 \\ -|x| + \frac{1}{2}, & \text{otherwise} \end{cases}, \tag{S48}$$

and

$$P^{FE}(\boldsymbol{\beta}) \propto \log\det(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}). \tag{S49}$$

Then, if $\tilde{\boldsymbol{\theta}}$ exists, $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$ is nondegenerate and all components of $\tilde{\boldsymbol{\beta}}$ are finite whenever there is a $\boldsymbol{\theta}$ in the interior of Θ such that $\ell(\boldsymbol{\theta}) > -\infty$.

Proof. We first note that both penalties $P^{FE}(\boldsymbol{\beta})$ and $P^{MVRE}(\boldsymbol{L})$ as well as the log-likelihood $\ell(\boldsymbol{\theta})$ are bounded from above. From (S48) it is immediate that (S47) is bounded from above by zero (for $\boldsymbol{L} = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix). Boundedness from above of $\log\det(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})$ is shown in Theorem 2 of Kosmidis and Firth (2020). Boundedness from above of $\ell(\boldsymbol{\theta})$ follows from the observation that the integral inside the log of $\ell(\boldsymbol{\theta})$ integrates a probability mass function, which always lies in $[0, 1]$ with respect to a normal density, so that $\ell(\boldsymbol{\theta})$ is bounded from above by zero. Next, we show that whenever $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is on the boundary of Θ, then either $P^{FE}(\boldsymbol{\beta}) = -\infty$ or $P^{MVRE}(\boldsymbol{L}) = -\infty$. For this, let $\Theta = \Theta_{FE} \cup \Theta_{MVRE}$, so that $\boldsymbol{\beta} \in \Theta_{FE}$ and $\boldsymbol{L} \in \Theta_{RE}$. Then $\boldsymbol{\theta} \in \partial\Theta$ if and only if $\boldsymbol{\beta} \in \partial\Theta_{FE}$ or $\boldsymbol{L} \in \partial\Theta_{MVRE}$, where $\partial\Theta_{FE}$ is the set of all $\boldsymbol{\beta}$ with at least one infinite component and $\partial\Theta_{MVRE}$ the set of all lower triangular matrices with nonnegative diagonal elements with at least one infinite component or at least one zero entry on the main diagonal. Kosmidis and Firth (2020) show that for any path $\boldsymbol{\beta}(r) \in \Theta_{FE}$ indexed by $r \in \Re$ such that $\lim_{r\to\infty} \boldsymbol{\beta}(r) = \boldsymbol{\beta}^\infty$, $\boldsymbol{\beta}^\infty \in \partial\Theta_{FE}$, $\lim_{r\to\infty} P^{FE}(\boldsymbol{\beta}(r)) = -\infty$. On the other hand, by definition of $\partial\Theta_{MVRE}$ and noting that $\lim_{|x|\to\infty} P^{RE}(x) = -\infty$, it follows that for any sequence of lower triangular matrices $\boldsymbol{L}(r) \in \Theta_{MVRE}$ such that $\lim_{r\to\infty} \boldsymbol{L}(r) = \boldsymbol{L}^\infty$, $\boldsymbol{L}^\infty \in \partial\Theta_{MVRE}$, $\lim_{r\to\infty} P^{MVRE}(\boldsymbol{L}(r)) = -\infty$. Hence for any sequence of $\boldsymbol{\theta}(r)$ such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) = \boldsymbol{\theta}^\infty$ for $\boldsymbol{\theta}^\infty \in \partial\Theta$, $\lim_{r\to\infty} \tilde{\ell}(\boldsymbol{\theta}(r)) = -\infty$. Now if there exists a $\boldsymbol{\theta}$ in the interior of Θ such that $\ell(\theta) > -\infty$, then $\ell(\tilde{\boldsymbol{\theta}}) > -\infty$ and we conclude that $\tilde{\boldsymbol{\theta}}$ cannot lie on the boundary. By Lemma S4.1, we then conclude that $\tilde{\boldsymbol{\Sigma}}$ is not degenerate. This concludes the proof. □

S5 Further Simulations

S5.1 Simulation 1: Extreme fixed effects

This section presents a simulation that seeks to provoke degenerate fixed effects estimates through a strong dependence of the responses on the particular fixed effects – a phenomenon known to occur in standard logistic regression (Albert and Anderson, 1984; Kosmidis and Firth, 2020).

For this, we simulate from a Bernoulli-response GLMM with univariate random effects and logistic link function as follows. For five clusters $i = 1, \ldots, 5$ and within cluster observations $j = 1, \ldots, n$, $n \in \{50, 100, 200\}$, we draw an i.i.d. vector of fixed effects covariates $\boldsymbol{x}_{ij} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ where $X_{i1} = 1, X_{i2} \sim \mathrm{N}(0, 1), X_{i3} \sim \mathrm{Ber}\left(\frac{1}{2}\right), X_{i4} \sim \mathrm{Ber}\left(\frac{1}{4}\right)$, and $X_{i5} \sim \exp(1)$. The fixed effect covariates are drawn once and held fixed over the simulation. To control the degree of dependence of the responses on a particular fixed effect covariate, the parameter of fixed effects is set as $\boldsymbol{\beta} = (1, -0.5, \lambda, 0.25, -1)$, where λ takes integer values from -10 to 10. For each specification of n, λ, we draw 100 samples from the model

$$\begin{aligned} Y_{ij} \mid u_i &\sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i, \\ u_i &\sim \mathrm{N}(0, 9) \quad (i = 1, \ldots, 5; j = 1, \ldots, n), \quad n \in \{50, 100, 200\} \end{aligned} \tag{S50}$$

The random effects dispersion parameter $\sigma = 3$ was chosen as to avoid estimation issues associated with small random effects. We estimate the parameters using our proposed MSPAL with the penalties given in Section 5 of the main text, MAL and `bglmer` from the `blme` R package (Chung et al., 2013) with a normal and t prior for the fixed effects and a gamma prior for the random effects variance. We approximate the log-likelihood with a 20-point adaptive Gauss-Hermite quadrature approximation. For MAL and MSPAL, we optimize the approximate log-likelihood using the optimization methods "CG", "BFGS", "nlminb" and "L-BFGS-B" from the `optimx` R package (Nash and Varadhan, 2011) and report the best fit. Both `bglmer` specifications use the default `bglmer` optimization settings.

Table S1 shows the number of estimates per specification which resulted in an degenerate parameter estimate. We considered an estimate degenerate, if it is larger than 50 in absolute value, the gradient is larger than 0.001 in absolute value, or if the estimated asymptotic standard errors are larger than 40. Figure S1 shows the dispersion of the estimates β_3 around the true value (indicated by dashed horizontal line) per specification for all estimation methods. For presentability, the y-axis has been cropped to omit overly extreme estimates. For the MAL, 672 estimates are cropped, while for `bglmer`(n), one estimation is not shown due to failed estimation (NaN output). Note that the boxplots for the MAL do not depict the empirical distribution of the MAL estimate of β_3 over different samples as theoretically infinite estimates assume finite values in estimation due to numerical precision limitations and resulting premature declaration of convergence. We see from Table S1 that the MSPAL gives the most stable parameter estimation, with one out of 6000 samples exhibiting estimation issues due to failed convergence. The MAL on the other hand becomes highly problematic for large values of β_3 and even returns degenerate estimates for moderate values of β_3 with non-negligible frequency. The `bglmer` estimates, even though they penalize fixed effects more harshly, as can be seen by the shrinkage-induced bias of the `bglmer` estimates of β_3 in Figure S1, they encounter estimation issues substantially more often than the MSPAL estimates.

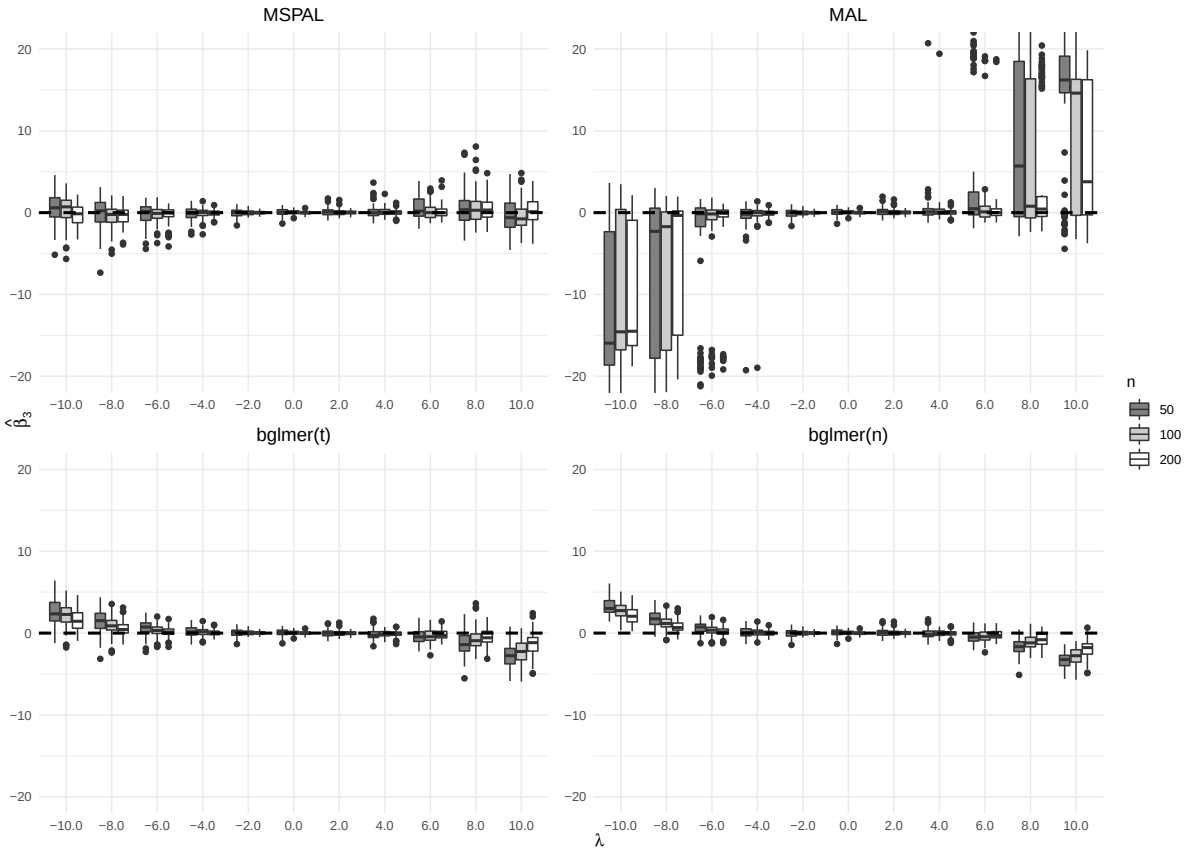

Figure S1: Centered estimation output of $\hat{\beta}_3 - \beta_3$ from Simulation 1

Table S1: Percentage of degenerate estimates from Simulation 1

		λ										
		-10	-8	-6	-4	-2	0	2	4	6	8	10
MSPAL	n=50	0	0	0	0	0	0	0	0	0	0	0
	n=100	0	0	0	0	0	0	0	0	0	**1**	0
	n=200	0	0	0	0	0	0	0	0	0	0	0
MAL	n=50	**73**	**46**	**16**	**2**	0	0	0	**1**	**15**	**49**	**80**
	n=100	**63**	**45**	**10**	**1**	0	0	0	**1**	**5**	**31**	**61**
	n=200	**65**	**26**	**8**	0	0	0	0	0	**2**	**24**	**49**
bglmer(t)	n=50	0	0	0	0	0	0	0	0	0	0	0
	n=100	0	0	0	**1**	0	0	0	0	0	0	0
	n=200	0	0	**1**	0	**1**	0	0	0	0	0	**1**
bglmer(n)	n=50	**1**	0	0	**1**	**1**	**1**	0	0	0	**1**	0
	n=100	0	**1**	**1**	0	**2**	0	**1**	**1**	**1**	0	0
	n=200	0	0	0	0	0	0	**2**	**1**	0	0	**3**

S5.2 Simulation 2: Extreme random effects variance

In this simulation, we seek to provoke degenerate random effects variance estimates, that is random effects variance estimates that are either zero or infinite. One of the peculiarities of Bernoulli-response (or Binomial-response) GLMMs is that there can be separation of the observations with respect to the random effects covariates. Analogously to separation in logistic regression models (Albert and Anderson, 1984), where covariate constellations such that the responses can be separated by a hyperplane spanned by the covariate column vectors, lead to infinite parameter estimates, it is known that certain constellations of random effects covariates can lead to data separation and consequently degenerate random effects estimates (see for example Sauter and Held (2016) or the discussion on `https://stats.stackexchange.com/questions/44755`). We consider a simple simulation to provoke such data configurations by simulating from a Bernoulli-response GLMM with univariate random effects and logistic link function and vary the dependence of the responses on the grouping variable by controlling the random effects variance parameter.

For five clusters $i = 1, \ldots, 5$ and within cluster observations $j = 1, \ldots, n$, $n \in \{50, 100, 200\}$, we draw an i.i.d. vector of fixed effects covariates $\boldsymbol{x}_{ij} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ where $X_{i1} = 1$, $X_{i2} \sim \mathrm{N}(0, 1)$, $X_{i3} \sim \mathrm{Ber}\left(\frac{1}{2}\right)$, $X_{i4} \sim \mathrm{Ber}\left(\frac{1}{4}\right)$, and $X_{i5} \sim \exp(1)$. The fixed effect covariates are drawn once and held fixed over the simulation. Likewise, the fixed effects $\boldsymbol{\beta} = (1, -0.5, 0.5, 0.25, -1)$ are held fixed over the simulation, while $\lambda = \log \sigma$ is varied over the integer values from -5 to 2. For each specification of n, λ, we draw 100 samples from the model

$$\begin{aligned} Y_{ij} \mid u_i &\sim \mathrm{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i, \\ u_i &\sim \mathrm{N}(0, \exp(\lambda)^2) \quad (i = 1, \ldots, 5; j = 1, \ldots, n), \quad n \in \{50, 100, 200\} \end{aligned} \tag{S51}$$

We estimate the parameters using our proposed MSPAL with the penalties given in Section 5 of the main text, MAL and `bglmer` from the `blme` R package (Chung et al., 2013) with a normal and t prior for the fixed effects and a gamma prior for the random effects variance. We approximate the log-likelihood with a 20-point adaptive Gauss-Hermite quadrature approximation. For MAL and MSPAL, we optimize the approximate log-likelihood using the optimization methods "CG", "BFGS", "nlminb" and "L-BFGS-B" from the `optimx` R package (Nash and Varadhan, 2011) and report the best fit. Both `bglmer` specifications use the default `bglmer` optimization settings. Figure S2 shows the dispersion of the estimates for $\log \sigma$ around the true value (indicated by dashed horizontal line), for each estimation method and specification of λ, n. For the MAL and `bglmer` estimates, these boxplots do not approximate the distribution of the maximum likelihood estimator as, owed to numerical precision limitations, parameter estimates which ought to be infinite or zero are not estimated as such so that the point masses at the boundaries of the parameter space are missing. For blgmer(t), 6 estimates are not shown and for bglmer(n), 10 estimates are not shown due to failed estimation. While both the MSPAL and `bglmer` shrink negative estimates of $\log \sigma$ towards zero, the shrinkage induced by `bglmer` is considerably stronger, as can be seen by the absolute amount of shrinkage and the smaller dispersion of the estimates. Moreover, we see that for larger values of $\log \sigma$ `bglmer` is unable to guard against infinite estimates of the random effects variance.

Table S2 shows the number of estimates for $\log \sigma$ per specification which resulted in an degenerate random effects variance estimate. We considered an estimate degenerate, if it is larger than $\log(50)$ or smaller than $\log(0.01)$, the gradient is larger than 0.001 in absolute value, or if the estimated asymptotic standard errors are larger than 40. We see that MSPAL is the most stable estimation routine and that both `bglmer` and MAL exhibit estimation degeneracies frequently for both small and large values of the random effects variance.

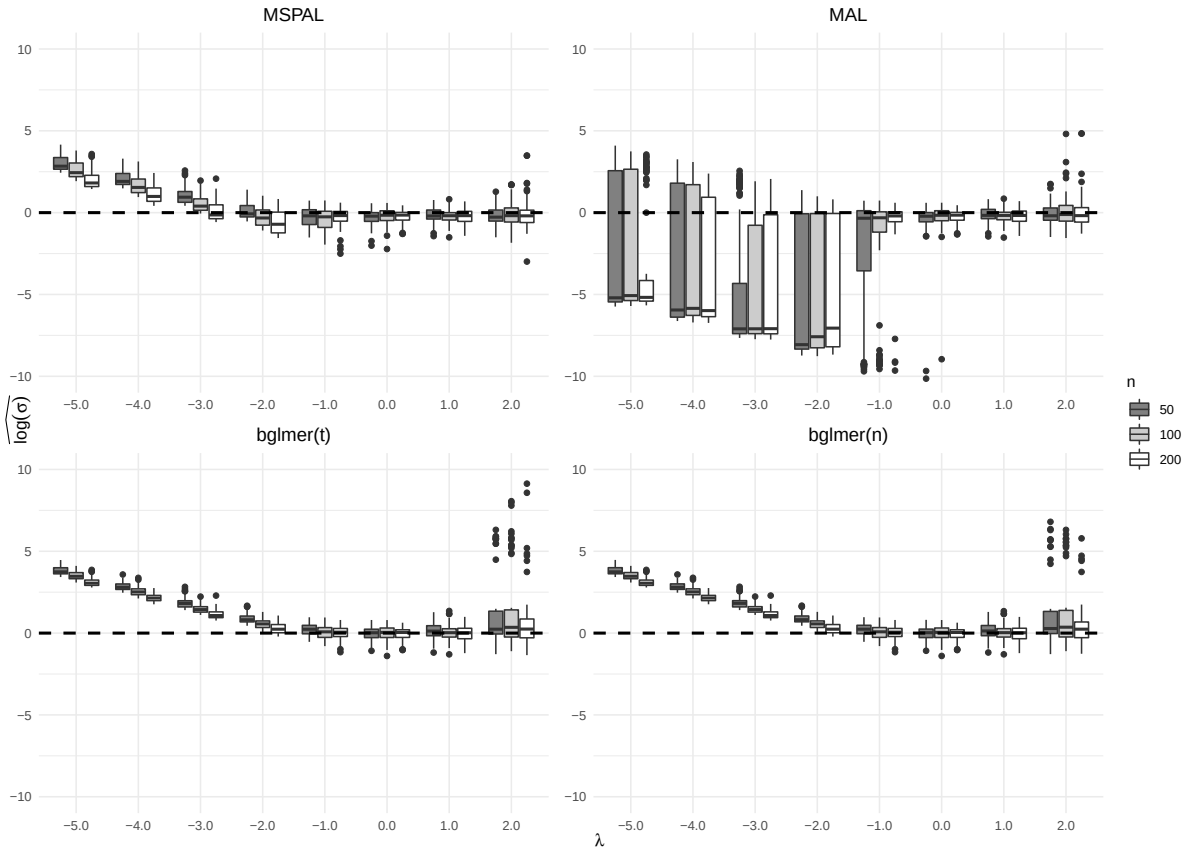

Figure S2: Centered estimation output of $\widehat{\log \sigma} - \log(\sigma)$ from Simulation 2

Table S2: Percentage of degenerate estimates from Simulation 2

		λ							
		-5	-4	-3	-2	-1	0	1	2
MSPAL	n=50	0	0	0	0	0	0	0	**1**
	n=100	0	0	0	0	0	0	0	0
	n=200	0	0	0	0	0	0	0	**2**
bglmer(t)	n=50	0	0	0	0	0	0	0	**12**
	n=100	**1**	**1**	**6**	**4**	**2**	0	0	**17**
	n=200	**25**	**31**	**19**	**24**	**3**	0	0	**11**
bglmer(n)	n=50	**4**	**1**	**4**	**3**	**2**	**1**	**1**	**13**
	n=100	**21**	**19**	**14**	**14**	**11**	**2**	**1**	**20**
	n=200	**33**	**45**	**40**	**28**	**23**	**4**	0	**17**
MAL	n=50	**73**	**70**	**76**	**70**	**25**	**2**	0	0
	n=100	**68**	**66**	**75**	**58**	**16**	**1**	0	**7**
	n=200	**78**	**68**	**69**	**52**	**4**	0	0	**5**

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Chung, Y., S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika 78*(4), 685–709.

Gilbert, P. and R. Varadhan (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1.

Harville, D. A. (1998). Matrix algebra from a statistician's perspective.

Kosmidis, I. and D. Firth (2020, 08). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika 108*(1), 71–82.

Magnus, J. R. and H. Neudecker (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

Nash, J. C. and R. Varadhan (2011). Unifying optimization algorithms to aid software system users: optimx for r. *Journal of Statistical Software 43*(9), 1–14.

Ogden, H. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika 104*(1), 153–164.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauter, R. and L. Held (2016). Quasi-complete separation in random effects of binary response mixed models. *Journal of Statistical Computation and Simulation 86*(14), 2781–2796.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.