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

Abstract

We introduce a soft penalization approach for stabilizing maximum likelihood estimation in Bernoulli-response generalized linear mixed models, which is known to have a positive probability to result in estimates on the boundary of the parameter space. Such estimates, instances of which are infinite values for fixed effects and singular or infinite variance components, can cause havoc to numerical estimation procedures and inference. We introduce an additive penalty to the log-likelihood function, which consists of appropriately scaled versions of the Jeffreys prior for the model with no random effects and the negative Huber loss. The resulting maximum softly-penalized likelihood estimates are shown to lie in the interior of the parameter space. Appropriate scaling of the penalty guarantees that the penalization is soft-enough to recover the optimal asymptotic properties expected by the maximum likelihood estimator, namely consistency, asymptotic normality, Cramér-Rao efficiency and asymptotically valid hypothesis testing. Further, our choice of penalties and scaling factor preserves invariance of the fixed effects estimates under linear transformation of the model parameters, such as contrasts. Maximum softly-penalized likelihood is compared to competing approaches on two real-data examples, and comprehensive simulation studies that illustrate its superior finite sample performance.

Keywords: logistic regression, infinite estimates, singular variance components, data separation, Jeffreys prior

1 Introduction

Generalized Linear Mixed Models (GLMMs; Charles E. McCulloch 2008, Chapter 7) are a potent class of statistical models that allow associating Gaussian and non-Gaussian responses, such as counts, proportions, positive responses, and so on, with covariates, while accounting for complex multivariate dependencies. This is achieved by linking the expectation of a response to a linear combination of covariates and parameters (fixed effects), and sources of extra variation (random effects) with known distributions. Although these models find application in numerous fields such as biology, ecology and the social sciences (Bolker et al., 2009), estimation of GLMMs is not straightforward in practice, because their likelihood is generally an intractable multivariate integral.

Maximum approximate likelihood (MAL) methods maximize an approximation of the GLMM likelihood, that can, in principle, be chosen to be arbitrarily accurate (see, for example, Raudenbush et al., 2000; Pinheiro and Chao, 2006). Such methods are pervasive in contemporary GLMM practice because, like maximum likelihood (ML), MAL estimators are consistent under general conditions about the model, and the MAL estimates and the approximate likelihood itself can be used for the construction of likelihood-based inferences, such as likelihood-ratio tests or Wald statistics, and can be used for model selection based on information criteria. An alternative approach to MAL are Bayesian posterior update procedures (see, for example, Zhao et al., 2006; Browne and Draper, 2006). However, they come with various technical difficulties, such as determining the scaling of the covariates, selecting appropriate priors, coming up with efficient posterior sampling algorithms, and determining burn-in times of chains for reliable estimation. Yet another alternative to MAL are maximum penalized quasi-likelihood (MPQL) methods (Schall, 1991; Wolfinger and O'connell, 1993; Breslow and Clayton, 1993) which essentially fit a Linear Mixed Model to transformed pseudo-responses. However, the penalized quasi likelihood may not yield an accurate approximation of the GLMM likelihood. As a result, MPQL estimators can have large bias when the random effects variances are large (Bolker et al., 2009; Rodriguez and Goldman, 1995) and are not necessarily consistent (Jiang, 2017, Chapter 3.1).

Despite the pervasiveness of MAL, certain data configurations can result in MAL estimates of the variance-covariance matrix of the random effects distribution to be on the boundary of the parameter space, such as infinite or zero estimated variances, or, more generally, singular estimates of the variance-covariance matrix; see Chung et al. (2013) for an excellent discussion. In addition, as is the case in maximum likelihood estimation of Bernoulli-response generalized linear models (GLMs; see, for example McCullagh and Nelder, 1989, Chapter 4), the MAL estimates of the fixed effects can be infinite. As is well-acknowledged in the GLMM literature (see, for example Bolker et al., 2009; Bolker, 2018; Pasch et al., 2013), both instances of boundary estimates can cause havoc to numerical optimization procedures used for MAL. In addition, if they go undetected, they can substantially impact first-order inferential procedures, like Wald tests, resulting in spuriously strong or weak conclusions. In contrast to the numerous approaches to detect (see, for example, Kosmidis and Schumacher 2021 for the `detectseparation` R package that implements the methods in Konis 2007) and handle (see, for example, Kosmidis and Firth, 2020; Heinze and Schemper, 2002; Gelman et al., 2008) infinite estimates in Bernoulli-response GLMs, little methodology or guidance is available on how to detect or deal with degenerate estimates in GLMMs.

We introduce a maximum softly-penalized approximate likelihood (MSPAL) procedure for Bernoulli-response GLMMs that returns estimators that are guaranteed to take values in the interior of the parameter space, and are also consistent, asymptotically normal, Cramér-Rao efficient and give asymptotically valid inference, under no additional assumptions beyond those typically required for establishing consistency, asymptotic normality, and asymptotically valid inference of MAL or ML estimators. Although the developments here are for Bernoulli-response GLMMs, they provide a blueprint for the construction of penalties and estimators with values in the interior of the parameter space for any GLMM and, more generally, for M-estimation settings where boundary estimates occur. The (approximate) likelihood penalty we introduce consists of appropriately scaled versions of the Jeffreys prior for the model with no random effects, and the negative Huber loss. We show that the MSPAL estimates are guaranteed to be in the interior of the parameter space, and impose a scaling to the penalty that guarantees that i) penalization is soft-enough for the MSPAL estimator to have the optimal asymptotic properties expected by the ML estimator, and ii) that the fixed effects estimates are invariant to linear transformation of the model parameters, such as contrasts, in the sense that the MSPAL estimates of linear transformations of the fixed effects parameters are the linear transformations of the MSPAL estimates. Both i) and ii) are in contrast to other penalization procedures that have been proposed in the literature (see, for example, Chung et al., 2013, 2015). Maximum softly-penalized likelihood is compared to prominent competing approaches through two real-data examples, and comprehensive simulation studies that illustrate its superior finite-sample performance.

The remainder of the paper is organized as follows. Section 2 defines the clustered Bernoulli-response GLMM and Section 3 gives a motivating real-data example of degenerate maximum approximate likelihood estimates in a Bernoulli-response GLMM. Section 4 formalizes the maximum softly penalized approximate likelihood framework and states the large sample results of the softly penalized maximum likelihood estimator. Section 6 demonstrates the performance of the MSPAL on another real-data example and a data-based simulation and Section 7 provides concluding remarks. Proofs, details on the simulations in this paper and further simulations on synthetic data are given in the supplementary material.

2 Bernoulli-response generalized linear mixed models

Suppose that response vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are observed with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top \in \{0,1\}^{n_i}$, possibly along with covariate matrices $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k$, respectively, where $\boldsymbol{V}_i$ is a $n_i \times s$ matrix.

A Bernoulli-response GLMM assumes that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are realizations of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, whose entries $Y_{i1}, \ldots, Y_{in_i}$ $(i = 1, \ldots, k)$ are independent Bernoulli random variables conditionally on a vector of random effects $\boldsymbol{u}_i$. The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ are assumed to be independent realizations of a multivariate normal distribution, and the conditional mean of each Bernoulli random variable is linked to a linear predictor η_{ij}, which is a linear combination of covariates with random effects and fixed effects. Specifically,

$$Y_{ij} \mid \boldsymbol{u}_i \sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i \tag{1}$$

$$\boldsymbol{u}_i \sim \text{N}(\boldsymbol{0}_q, \boldsymbol{\Sigma}) \quad (i = 1, \ldots, k; j = 1, \ldots, n_i)\,, \tag{2}$$

where $\mu_{ij} = P(Y_{ij} = 1 \mid \boldsymbol{u}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, and $g : (0,1) \to \Re$ is a known monotone increasing link function, like the logistic, probit or complementary log-log. The vector $\boldsymbol{x}_{ij}$ is the jth row of the $n_i \times p$ model matrix $\boldsymbol{X}_i$ associated with the p-vector of fixed effects $\boldsymbol{\beta} \in \Re^p$, and $\boldsymbol{z}_{ij}$ is the jth row of the $n_i \times q$ model matrix $\boldsymbol{Z}_i$ associated with the q-vector of random effects $\boldsymbol{u}_i$. The model matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are formed from subsets of columns of $\boldsymbol{V}_i$. The variance-covariance matrix $\boldsymbol{\Sigma}$ collects the variance components and is assumed to be symmetric and positive definite. The Bernoulli-response GLMM in (1) is here introduced explicitly in terms of clusters. The other often encountered formulation of GLMMs in the literature (Charles E. McCulloch, 2008, Chapter 7.4) absorbs the clustering into the variance components structure of $\boldsymbol{\Sigma}$ and is therefore a clustered GLMM with a single cluster. Hence, the presentation and results here are with no loss of generality.

The marginal likelihood about $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ for model (1) is

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-kq/2} \det(\boldsymbol{\Sigma})^{-k/2} \prod_{i=1}^{k} \int_{\Re^q} \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1 - y_{ij}} \exp\left\{ -\frac{\boldsymbol{u}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_i}{2} \right\} d\boldsymbol{u}_i\,, \tag{3}$$

Formally, the ML estimator is the maximizer of (3) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. However, (3) involves intractable integrals, which are typically approximated before maximization, resulting in MAL estimators. For example, the popular `glmer` routine of the R (R Core Team, 2020) package `lme4` (Bates et al., 2015) uses adaptive Gauss-Hermite quadrature for one-dimensional random effects and Laplace approximation for higher-dimensional random effects. A detailed account of those approximation methods can be found in Pinheiro and Bates (1995).

3 Motivating example

The working data set in this section is a reduced version of the data in McKeon et al. (2012), as provided in the worked examples of Bolker (2015) (available at `https://bbolker.github.`

Table 1: Culcita data (McKeon et al., 2012) from the worked examples of Bolker (2015) (available at `https://bbolker.github.io/mixedmodels-misc/ecostats_chap.html`).

	Block									
Treatment	1	2	3	4	5	6	7	8	9	10
none	0,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,0
crabs	0,0	0,0	0,0	0,0	1,1	1,1	1,1	1,1	1,1	1,1
shrimp	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1	1,1
both	0,0	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1

`io/mixedmodels-misc/ecostats_chap.html`). The data is given in Table 1 and comes from trials involving coral-eating sea stars Culcita novauguineae (hereafter Culcita) attacking coral that harbour differing combinations of protective symbionts, involving crabs and shrimp. The design is a randomised complete block design with two replications per treatment per block, four treatments, involving no symbionts, crabs only, shrimp only, both crabs and shrimp, and ten temporal blocks. As a result there is a total of 80 observations on whether predation was present (recorded as one) or not (recorded as zero). By mere inspection of Table 1, we note that predation becomes more prevalent with increasing block number, and that predation gets suppressed when either crabs or shrimp are present, and more so when both symbionts are present. The only observation that deviates from this general trend is the observation in block 10 with no predation and no symbionts.

A Bernoulli-response GLMM with one random intercept per block can be used here to associate predation to treatment effects while accounting for heterogeneity between blocks. Such a model can be defined as

$$Y_{ij} \mid u_i \sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = \beta_0 + u_i + \beta_j \tag{4}$$

$$u_i \sim \text{N}(0, \sigma^2) \quad (i = 1, \ldots, 10; j = 1, \ldots, 4)\,. \tag{5}$$

In the above expressions, Y_{i1}, Y_{i2}, Y_{i3}, and Y_{i4} correspond to responses for "none", "crabs", "shrimp", "both", respectively, and we set $\beta_1 = 0$ for identifiability purposes, effectively using "none" as a reference category. The logarithm of the model likelihood (3) about the parameters $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_4)^\top$ and $\psi = \log \sigma$ for model (4) is approximated using an adaptive quadrature rule with $Q = 100$ points (see, for example, Liu and Pierce, 1994; Pinheiro and Bates, 1995) as implemented in `glmer`. The approximation to the log-likelihood gets more accurate as the number of quadrature points increases, and here we choose $Q = 100$ points, which is the maximum possible in the current `glmer` implementation.

All parameter estimates of model (4) reported in the current example are computed after removing the atypical observation with zero predation in block 10 when there are no symbionts. Estimates based on all data points are provided in Table **??** of the supplementary materials **#IK: include that**.

The MAL estimates of $\boldsymbol{\beta}$ and ψ in Table 2 are computed using the numerical optimization procedures "BFGS" and "CG" (MAL(BFGS) and MAL(CG), respectively), as these are readily available from the `optimx` R package (see Nash and Varadhan, 2011, Section 3 for details), with default starting values. The MAL(BFGS) and MAL(CG) estimates are different, and are notably extreme on the logistic scale. This is due to the two optimization procedures stopping early at different points in the parameter space after having prematurely declared convergence. The large estimated standard errors are indicative of the approximation to the log-likelihood being almost flat around the estimates. In this case, the MAL estimates for the fixed effects $\beta_0, \beta_1, \beta_2, \beta_3$ are in reality infinite in absolute value.

Table 2: Estimates from the degenerate Culcita subdataset of Bolker (2015) using MAL,MSPAL and `bglmer`

	MAL(BFGS)	MAL(CG)	bglmer(t)	bglmer(n)	MSPAL
		reference category: “none”			
β_0	15.88	15.37	6.39	4.90	8.41
	(10.14)	(9.50)	(2.60)	(2.08)	(3.43)
β_2	-12.93	-12.46	-4.02	-2.84	-7.22
	(9.15)	(8.53)	(1.59)	(1.27)	(3.21)
β_3	-14.81	-14.30	-4.81	-3.44	-8.26
	(9.89)	(9.24)	(1.73)	(1.35)	(3.48)
β_4	-17.71	-17.15	-6.47	-4.73	-10.10
	(10.70)	(10.02)	(2.05)	(1.57)	(3.84)
$\log\sigma$	2.31	2.28	1.72	1.54	1.80
	(0.64)	(0.62)	(0.44)	(0.43)	(0.45)
		reference category: “both”			
γ_0	-1.82	-1.74	0.37	0.57	-1.70
	(3.92)	(3.77)	(2.24)	(2.07)	(2.46)
γ_1	17.74	17.09	6.70	5.75	10.10
	(10.75)	(10.03)	(2.19)	(1.88)	(3.84)
γ_2	4.78	4.65	1.63	1.26	2.88
	(3.08)	(2.98)	(1.43)	(1.32)	(1.85)
γ_3	2.89	2.83	0.83	0.56	1.85
	(2.27)	(2.22)	(1.35)	(1.28)	(1.60)
$\log\sigma$	2.31	2.28	1.74	1.66	1.80
	(0.64)	(0.62)	(0.44)	(0.44)	(0.45)

Parameter estimates are also obtained using the `bglmer` routine of the `blme` R package (Chung et al., 2013) that has been developed to ensure that parameter estimates from GLMMs are away from the boundary of the parameter space. The estimates shown in Table 2 are obtained using a penalty for σ inspired by a gamma prior (default in `bglmer`; see Chung et al. 2013 for details) and two of the default prior specifications for the fixed effects: i) independent normal priors (“bglmer(n)”), and ii) independent t priors (“bglmer(t)”), as these are implemented in `blme`; see `bmerDist-class` in the help pages of `blme` for details. We also show the estimates obtained using the MSPAL estimation method that we propose in the current work.

The maximum penalized approximate likelihood estimates from `bglmer` and the corresponding estimated standard errors appear to be finite. Nevertheless, the use of the default priors directly breaks parametrization invariance under contrasts, which MAL estimates enjoy. For example, Table 2 also shows the estimates of model (4) with $\eta_{ij} = \gamma_0 + u_i + \gamma_j$, where $\gamma_4 = 0$, i.e. setting “both” as a reference category. Hence, the identities $\gamma_0 = \beta_0 + \beta_4$, $\gamma_1 = -\beta_4$, $\gamma_2 = \beta_2 - \beta_4$, $\gamma_3 = \beta_3 - \beta_4$ hold, and it is natural to expect those identities from the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. As is evident from Table 2, the `bglmer` estimates with either normal or t priors can deviate substantially from those identities. For example, the `bglmer` estimate of γ_1 based on normal priors is 5.75 while that for β_4 is -4.73, and the estimate of $\log\sigma$ is 1.54 in the $\boldsymbol{\beta}$ parametrization and 1.66 in the $\boldsymbol{\gamma}$ parametrization. Furthermore, different contrasts give varying amounts of deviations from these identities. On the other hand, the approximate likelihood is invariant to monotone parameter transformations. As a result, the corresponding identities hold exactly for the MAL estimates with the deviations observed in Table 2 being due to early

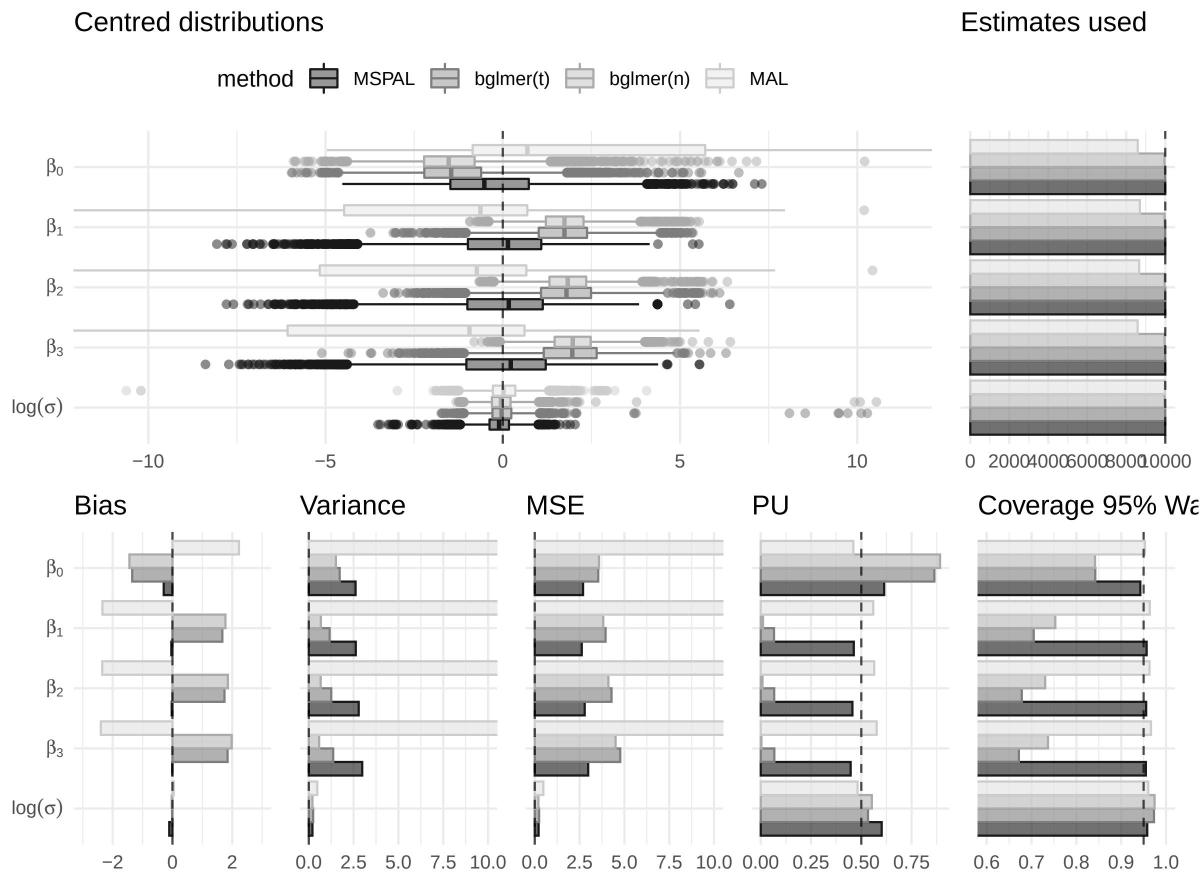

Figure 1: Performance metrics for parameter estimates of MAL,MSPAL and `bglmer` from simulating a Bernoulli-response GLMM from the Culcita data at the MAL

stopping of the optimization routines.

The `bglmer` estimates are typically closer to zero in absolute value than the MAL estimates because the normal and t priors are all centred at zero. Furthermore, the estimates using normal priors tend to shrink more towards zero than those using t priors, because the latter have heavier tails than the former. In order to assess the impact of shrinkage on the frequentist properties of the estimators, we simulate 10000 independent samples of responses for the randomized complete block design in Table 1, at the MAL estimates in the $\boldsymbol{\beta}$ parametrization when all data points are used (see Table **??** of the Supplementary Materials **#IK: include that**). For each sample, we compute the MAL and MSPAL estimates, as well as the `bglmer` estimates based on normal and t priors.

Figure 1 shows boxplots for the sampling distributions of the estimators, centred at the true value, the estimated finite-sample bias, variance, mean squared error, and probability of underestimation for each estimator, along with the estimated coverage of 95% Wald confidence intervals based on the estimates and estimated standard errors from the negative Hessian of the approximate log-likelihood at the estimates. The plotting range for the support of the distributions has been restricted to $(-11, 11)$, which does not contain all MAL estimates in the simulation study but contains all estimates for the other methods. We should note here that apart from the estimated probability of underestimation, estimates for the other summaries are not well-defined for MAL, because the probability of boundary estimates is positive. In fact, there were issues with at least one of the MAL estimates for 9.25% of the simulated samples.

These issues are either due to convergence failures or because the estimates or estimated standard errors have been found to be atypically large in absolute value. The displayed summaries for MAL are computed based only on estimates which have not been found to be problematic. Clearly, the amount of shrinkage induced by the normal and t priors is excessive. Although the resulting estimators have small finite-sample variance (with the one based on normal priors having the smallest), they have excessive finite-sample bias, which is often at the order of the standard deviation resulting in large mean squared errors, and the sampling distributions to be located far from the respective true values. Importantly, the combination of small variance and large bias readily impacts first-order inferences; Wald-type confidence intervals about the fixed effects are found to systematically undercover the true parameter value. Finally, both bglmer(n) and bglmer(t) do not appear prevent extreme positive variance estimates.

As is apparent from Table 2, the identities on the model parameters hold exactly with the proposed MSPAL estimates, where the observed deviations are attributed to rounding errors. Furthermore, from Figure 1 we see that the penalty we propose not only ensures that estimates are away from the boundary of the parameter space, but its soft nature guarantees that estimators have the optimal frequentist properties that would be expected by the MAL estimator had it not taken boundary values.

4 Penalized likelihoods

4.1 Setup

Suppose that we observe the values $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ of a sequence of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$ with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top \in \mathcal{Y} \subset \Re^{n_i}$, possibly with a sequence of covariate vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, with $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{is})^\top \in \mathcal{X} \subset \Re^s$. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_k^\top)^\top$, and denote by $\boldsymbol{V}$ the set of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Further, assume that the data generating process of $\boldsymbol{Y}$, conditional on $\boldsymbol{V}$ has a density or probability mass function $f(\boldsymbol{Y} \mid \boldsymbol{V}; \boldsymbol{\theta})$, indexed by a parameter $\boldsymbol{\theta} \in \Theta \subset \Re^d$. Denote the parameter that identifies the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{V}$ by $\boldsymbol{\theta}_0 \in \Theta$.

A popular method for estimating the parameter vector $\boldsymbol{\theta}_0$ is to maximize the logarithm of the likelihood $f(\boldsymbol{Y} \mid \boldsymbol{V}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. If the likelihood is not available in closed form then an approximation of it may be maximized instead. For example, expression (3) gives $f(\boldsymbol{Y} \mid \boldsymbol{V}; \boldsymbol{\theta})$ in the case of the Bernoulli-response GLMMs of Section 2, and in Section 3 we use an adaptive Gauss-Hermite quadrature approximation to the log-likelihood. In what follows, $\ell(\boldsymbol{\theta})$ denotes either the log-likelihood or an approximation to it whenever that distinction is immaterial for the context. Furthermore, the dependence of $\ell(\boldsymbol{\theta})$ on $\boldsymbol{Y}$ and $\boldsymbol{V}$ is suppressed for notational convenience. Then, the ML (or MAL) estimator of $\boldsymbol{\theta}$ is defined as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta})$.

Let $\tilde{\boldsymbol{\theta}}$ be the the maximum penalized likelihood (MPL) (or maximum penalized approximate likelihood; MPAL) estimator

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \{\ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta})\},$$

where $P(\boldsymbol{\theta})$ is an additive penalty to $\ell(\boldsymbol{\theta})$ that may depend on $\boldsymbol{Y}$ and $\boldsymbol{V}$.

In the remainder of this section we derive the conditions that $P(\boldsymbol{\theta})$ must satisfy to ensure that the MPL or MPAL estimator $\tilde{\boldsymbol{\theta}}$ i) takes values always in the interior of Θ, ii) is invariant under linear transformations of the parameters, such as scaled contrasts that are usually employed with categorical covariates in regression modelling, and iii) has similar first order asymptotics to $\hat{\boldsymbol{\theta}}$. We then derive a penalty that satisfies those conditions for Bernoulli-response GLMMs.

4.2 Interior point parameter estimates

Denote by $\partial\Theta$ the boundary of the parameter space, and let $\boldsymbol{\theta}(r)$, $r \in \Re$, be a path in the parameter space such that $\lim_{r \to \infty} \boldsymbol{\theta}(r) \in \partial\Theta$. A common approach to resolving issues with ML or MAL estimates being in $\partial\Theta$, like those encountered in the example of Section 3, is to instead use MPL or MPAL estimators from a penalty that satisfies $\lim_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$ and which are bounded from above. Then, if there is at least one point $\boldsymbol{\theta} \in \Theta$ such that $\ell(\boldsymbol{\theta}) > -\infty$, it must hold that $\tilde{\boldsymbol{\theta}}$ is in the interior of Θ.

For example, the penalties arising from the independent normal and independent t prior structures implemented in `blme` are such that $\lim_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$, whenever $\boldsymbol{\theta}(r)$ diverges to the boundary of the parameter space for the fixed effects. As a result, the bglmer(n) and bglmer(t) estimates for the fixed effects in Table 2) are finite. On the other hand, the default gamma-prior like penalty used in `bglmer` for the variance component σ is $-1.5 \log \sigma$, which, while it ensures that the estimate of $\log \sigma$ is not minus infinity, does not guard from positive infinite estimates. This is apparent in Figure 1, where several extreme positive bglmer(n) and bglmer(t) estimates are observed for $\log \sigma$; see, also, the vignettes of the `glmmsr` (Ogden, 2019) R package for an example with infinite variance component estimate in a Bernoulli-response GLMM.

4.3 Invariance under scaled linear transformations

The ML estimates are known to be invariant to transformations of the model parameters (see, for example Zehna, 1966). A particularly useful class of transformations in regression modelling with categorical covariates is the collection of scaled linear transformations $\boldsymbol{\theta}' = \boldsymbol{C}\boldsymbol{\theta}$ for known, invertible, real matrices $\boldsymbol{C}$. With such transformations one can obtain ML or MAL estimates and corresponding estimated standard errors for arbitrary sets of scaled parameter contrasts, when estimates for one of those sets of contrasts are available and with no need to re-estimate the model. Further, these transformations eliminate estimation and inferential ambiguity when two independent researchers analyse the same data set using the same model but with different contrasts, e.g. due to software defaults.

The example in Section 3 shows that not all MPL or MPAL estimators are invariant to linear transformations of the parameters. The condition required for achieving invariance is that the penalty satisfies $P(\boldsymbol{C}\boldsymbol{\theta}) = P(\boldsymbol{\theta}) + b$, where $b \in \Re$ is a real constant. This requirement does not hold for the penalties arising from the normal and t prior structures that are used to compute the bglmer(n) and bglmer(t) fixed effect estimates in Table 2. Hence, the bglmer(n) and bglmer(t) MPAL estimates are not invariant under linear transformations of the parameters.

4.4 Asymptotic properties

Consistency, asymptotic normality and valid asymptotic hypothesis testing of the proposed MSPAL estimator follow readily from similar such results for MAL estimators where the approximation error to the model log-likelihood is an additive error term. Indeed, the results presented in this section are a direct translation of the work of Ogden (2017), where the term "approximation error" is replaced by "penalty function". To state the results and their underlying assumptions, we introduce some further notation. Proofs are given in Section S2 of the supplementary material.

Let $S(\boldsymbol{\theta})$ be the score function of $\ell(\boldsymbol{\theta})$, i.e. $S(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta})$, and let $\tilde{S}(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}) + \nabla P(\boldsymbol{\theta})$ be the score of its penalized analogue $\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta})$. Denote the observed information matrix by $J(\boldsymbol{\theta}) = -\nabla\nabla^\top \ell(\boldsymbol{\theta})$. It is assumed that information regarding the model parameter accumulates at a rate r_n in the sense that $r_n^{-1} J(\boldsymbol{\theta}) \xrightarrow{p} I(\boldsymbol{\theta})$ as $n \to \infty$ for some nonrandom, positive definite, $\mathcal{O}(1)$, matrix $I(\boldsymbol{\theta})$ and with respect to some matrix norm $|||\cdot|||$. Further, let $\delta(\boldsymbol{\theta}) = \|\nabla P(\boldsymbol{\theta})\|$ for some vector norm $\|\cdot\|$, and for $S \subseteq \Theta$ define $\delta^\infty(S) = \sup_{\boldsymbol{\theta}\in S} \delta(\boldsymbol{\theta})$ and $\delta^\infty = \delta^\infty(\Theta)$. Finally, denote by $B_t(\boldsymbol{\theta})$ the ball of radius t around $\boldsymbol{\theta}$.

We impose standard M-estimation regularity conditions on the score function to establish consistency of $\tilde{\boldsymbol{\theta}}$ (see for example Vaart (1998, Chapter 5)).

A0 Both $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ are differentiable, with derivatives $S(\boldsymbol{\theta}), \tilde{S}(\boldsymbol{\theta})$

A1 $\sup_{\boldsymbol{\theta}\in\Theta} \left\| r_n^{-1} S(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta}) \right\| \xrightarrow{p} 0$ for some deterministic function $S_0(\boldsymbol{\theta})$

A2 For all $\varepsilon > 0$, $\inf_{\boldsymbol{\theta}\in\Theta: \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\varepsilon} \|S_0(\boldsymbol{\theta})\| > 0 = \|S_0(\boldsymbol{\theta}_0)\|$

A3 $\hat{\boldsymbol{\theta}}$ 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 4.1 (Consistency)**:** *Let $\delta^\infty = o_p(r_n)$, and assume that A0-A3 hold. Then $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.*
The regularity conditions we impose to establish asymptotic normality of $\tilde{\boldsymbol{\theta}}$ are standard conditions in maximum likelihood estimation.

A4 Both $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ are three times differentiable

A5 $\sup_{\boldsymbol{\theta}\in\Theta} \left|\left|\left| r_n^{-1} J(\boldsymbol{\theta}) - I(\boldsymbol{\theta}) \right|\right|\right| \xrightarrow{p} 0$ for some positive definite, nonrandom, $\mathcal{O}(1)$ matrix $I(\boldsymbol{\theta})$, that is continuous in $\boldsymbol{\theta}$ in a neighbourhood around $\boldsymbol{\theta}_0$

A6 $r_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, I(\boldsymbol{\theta}_0)^{-1})$

A7 $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$

Theorem 4.2 (Asymptotic Normality)**:** *Assume that conditions A3-A7 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 $r_n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, I(\boldsymbol{\theta}_0)^{-1})$.*

To state conditions for valid hypothesis testing using the MSPAL, let $\gamma^\infty(S) = \sup_{\boldsymbol{\theta}\in S} |||\nabla\nabla^\top P(\boldsymbol{\theta})|||$ and suppose we want to test $H_0 : \boldsymbol{\theta} \in \Theta^R$, where $\Theta^R \subset \Theta$ and $\dim(\Theta^R) < \dim(\Theta)$. Finally, let $\Lambda = 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}^R))$ and similarly $\tilde{\Lambda} = 2(\tilde{\ell}(\tilde{\boldsymbol{\theta}}) - \tilde{\ell}(\tilde{\boldsymbol{\theta}}^R))$, where $\hat{\boldsymbol{\theta}}^R, \tilde{\boldsymbol{\theta}}^R$ denote the maximizers of $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ over Θ^R respectively.

Theorem 4.3 (Hypothesis testing)**:** *Assume that conditions A3-A7 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 $H_0 : \tilde{\Lambda} - \Lambda = o_p(1)$.*

The conditions of Theorems 4.1-4.3 are one of many and other standard arguments to establish consistency and asymptotic of a maximum likelihood estimator are expected to lead to the same results. We note that the large sample results of the MSPAL operate under the assumption that $\ell(\boldsymbol{\theta})$ is the exact model likelihood or that $\ell(\boldsymbol{\theta})$ is an approximate likelihood for which the convergence and regularity assumptions of A0-A7 are with respect to a quantity of interest. Corollaries S2.1-S2.3 of the supplementary material give sufficient conditions about the approximation error to achieve the asymptotic results of Theorems 4.1-4.3 with an approximate likelihood. It is left to future research to link the approximation error rates of various approximation methods with these conditions. There are results on approximation errors of the log-likelihhod, that can be adapted to match our conditions. We refer the reader to Ogden (2021) for approximation errors to the log-likelihood in clustered GLMMs using Laplace's method, Ogden (2017) for approximation errors to the gradient of the log-likelihood with an example for an intercept-only Bernoulli-response GLMM, Stringer and Bilodeau (2022) for approximation errors to the log-likelihood in clustered GLMMs using Adaptive Gauss-Hermite quadrature and Jin and Andersson (2020) for general approximation errors for adaptive Gauss-Hermite quadrature.

4.5 Soft penalization

The conditions that we imposed on the penalty function for the asymptotic results of $\tilde{\boldsymbol{\theta}}$, namely $\delta^\infty = o_p(r_n)$ for consistency, and additionally $\delta^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n^{1/2})$ for some $t > 0$ for asymptotic normality and $\gamma^\infty(B_t(\boldsymbol{\theta}_0)) = o_p(r_n)$ for hypothesis testing, can be decomposed into a (uniform) boundedness condition on the gradient and Hessian of the penalty and a rate requirement. Hence, the following blueprint provides a straightforward way of constructing appropriate penalty functions. i) Find an unscaled penalty function $P_u(\boldsymbol{\theta})$ that guarantees estimates in the interior of the parameter space (see Section 4.2), ii) determine uniform bounds of $\|\nabla P_u(\boldsymbol{\theta})\|$, $|||\nabla\nabla^\top P_u(\boldsymbol{\theta})|||$ over Θ, and iii) rescale the penalty function in dependence of r_n to meet the rate requirements of Theorems 4.1-4.3. Note that the normal and t priors as well as the gamma and wishart priors that `bglmer` uses to penalize the fixed effects and variance components of a GLMM are not directly applicable in this framework as they do not have uniformly bounded gradients.

5 Softly-penalized likelihood for Bernoulli-response GLMMs

5.1 Fixed effects penalty

The unscaled fixed effects penalty we consider in this paper is the logarithm of Jeffreys invariant prior from a logistic GLM, that is

$$P_u^{FE}(\boldsymbol{\beta}) = \frac{1}{2}\log\det(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}).$$

Here $\boldsymbol{X}$ is the matrix of all fixed effect covariates, $\boldsymbol{W}$ is a diagonal matrix with diagonal entries $\boldsymbol{W}_{ii} = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$ and $\mu_i(\boldsymbol{\beta})$ is the inverse logit-transform of the fixed effects component of the linear predictor at a point $\boldsymbol{\beta}$ in the parameter space. For notational convenience, the dependence of $\mu(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$ is henceforth suppressed. Kosmidis and Firth (2020, Theorem 1) have shown that whenever $\boldsymbol{X}$ is full rank, then for any path $\boldsymbol{\beta}(r) \in \Re^p$ indexed by $r \in \Re$ such that $\lim_{r\to\infty} \boldsymbol{\beta}(r) = \boldsymbol{\beta}^\infty$, where $\boldsymbol{\beta}^\infty$ is an arbitrary point in $\Re^p$ with at least one infinite component, $\lim_{r\to\infty} \det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}) = 0$. Therefore, noting that (3) is always bounded from above by one as the conditional distribution of the response is a probability mass function, and that $\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})$ is nonzero for $\boldsymbol{\beta} = \mathbf{0}_p$ when $\boldsymbol{X}$ has full rank, adding this penalty to the log-likelihood guarantees finite fixed effect parameter estimates as long as there is one $\boldsymbol{\theta} \in \Theta$ for which the log-likelihood is not $-\infty$. Kosmidis and Firth (2020) show further, that Jeffreys invariant prior guarantees finite fixed effects estimates for other link functions, such as the probit, complementary log-log, log-log and cauchit link, so that the proposed penalty can be generalized to GLMMs where other link functions appear more natural.

The bounds on the first and second order partial derivatives of Jeffreys invariant prior in (6) and (7), can be used to establish the range of scaling factors that are in line with Theorems 4.1-4.3. In particular, we show in Theorem S3.1 of the supplementary material, that for any full rank matrix $\boldsymbol{X} \in \Re^{n\times p}$ and any $\boldsymbol{\beta} \in \Re^p$ it holds that

$$\left|\frac{\partial}{\partial\beta_i}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})\right| \le p \max_{1\le j\le n}|x_{ji}| \tag{6}$$

$$\left|\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})\right| \le 2p \max_{1\le k\le n}|x_{ki}| \max_{1\le k\le n}|x_{kj}| \tag{7}$$

Hence, as long as r_n is increasing n, any scaling factor that is $\mathcal{O}_p(\max_{i,j}|x_{ji}|^{-1})$ achieves appropriate scaling of Jeffreys invariant prior for consistency and asymptotic normality and any scaling that is $\mathcal{O}_p(\max_{i,j}|x_{ji}|^{-2})$ achieves valid asymptotic hypothesis testing.

We propose scaling Jeffreys invariant prior by $2\sqrt{p/n}$, which gives the scaled fixed effects penalty

$$P^{FE}(\boldsymbol{\beta}) = \sqrt{p/n}\log\det(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}) \tag{8}$$

By (6) and (7), it then follows that (8) is a valid penalty whenever $\max_{i,j} |x_{ji}| = \mathcal{O}_p(n^{1/2})$ as long as r_n is increasing in n. This certainly holds for bounded covariates, as considered in our real-data examples, as well as, for example, for covariate matrices whose entries are subgaussian random variables with common variance proxy σ^2, in which case $\max\limits_{i,j} |x_{ji}| = \mathcal{O}_p(\sqrt{2\sigma^2 \log(2np)})$ (see for example Rigollet (2015, Theorem 1.14)).

5.2 Variance components penalty

The variance components penalty we propose in this paper is the negative Huber loss function, and a multivariate generalization thereof, that is scaled appropriately to ensure asymptotic negligibility in line with Theorems 4.1-4.3.

We first consider the case univariate random effects, for which we propose to penalize $\log \sigma$ by the negative Huber loss with δ-parameter equal to one, that is

$$P_u^{RE}(\log \sigma) = \begin{cases} -\frac{1}{2}\{\log \sigma\}^2, & \text{if } |\log \sigma| \leq 1 \\ -|\log \sigma| + \frac{1}{2}, & \text{otherwise} \end{cases} \tag{9}$$

Following the discussion of Section 4.5, the variance components penalty of (9) must satisfy $\lim\limits_{\sigma \to 0} P_u^{\mathrm{RE}}(\sigma) = -\infty$ and $\sup_{\sigma \in (0,\infty)} \left\|\nabla P_u^{\mathrm{RE}}(\sigma)\right\|$ must be bounded. Note however that the domain of $P_u^{\mathrm{RE}}(\sigma)$, is bounded from below, so that if a penalty function $P_u^{\mathrm{RE}}(\sigma) : \Re_{>0} \to \Re$ is differentiable with uniformly bounded derivative over its domain, then it cannot be that $\lim\limits_{\sigma \to 0} P_u^{\mathrm{RE}}(x) = -\infty$. In the absence of a uniform bound on the gradient of the variance components penalty, it is not possible to apply the developed methodology to a penalty on the random effects variance parameter σ directly. A workaround is to parametrize the model in terms of $\log \sigma$, the range of which is $\Re$, rather than σ itself. For this reparametrized model, it is easily verified that the Huber loss of (9) has uniformly bounded first and second derivatives. Naturally, this implies that assumptions A0-A7 must apply to the reparametrized model. The continuous mapping theorem (see for example Vaart (1998, Theorem. 2.3)) and the delta method (see for example Vaart (1998, Chapter 3)) provide asymptotic results for the σ parametrization.

We propose scaling the negative Huber loss penalty by $2\sqrt{p/n}$ yielding the random effects penalty

$$P^{RE}(\log \sigma) = \sqrt{p/n} \begin{cases} -\{\log \sigma\}^2, & \text{if } |\log \sigma| \leq 1 \\ -2|\log \sigma| + 1, & \text{otherwise} \end{cases} \tag{10}$$

The negative Huber loss penalty on the log-transformed random effects variance can easily be extended to multivariate random effects. For this, we consider the Cholesky factorization, call it $\boldsymbol{L}$, of the variance components matrix $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. Since for positive definite matrices, the map from $\boldsymbol{\Sigma}$ to $\boldsymbol{L}$ is bijective, this reparametrization is well defined. To ensure that the diagonal entries of $\boldsymbol{L}$ are finite and positive, we penalize the logarithm of each main-diagonal entry by (10). To ensure finiteness of all lower-triangular entries off the main-diagonal, each entry is again penalized by the same penalty without the prior log-transform. This ensures that the resulting variance-covariance estimate, $\widetilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$, where all main-diagonal entries are transformed back to their natural parametrization, is nondegenerate. That is to say, $\boldsymbol{\Sigma}$ is symmetric, positive definite, with finite entries and exhibits no perfect estimated correlation, i.e. for all $i \neq j$, $\left|\frac{\widetilde{\boldsymbol{\Sigma}}_{ij}}{\sqrt{\widetilde{\boldsymbol{\Sigma}}_{ii}\widetilde{\boldsymbol{\Sigma}}_{jj}}}\right| < 1$. A proof is given in Lemma S4.1 of the supplementary material. Again, we require that all model regularity assumptions apply with the respect to log-transformed diagonal entries of $\boldsymbol{L}$, rather than $\boldsymbol{L}$. Large sample theory for $\widetilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$ follows from the continuous mapping theorem and the delta method.

The Theorem below establishes that our proposed penalties give estimates in the interior of the parameter space for a Bernoulli-response GLMM. A proof is given in Section 4.2 of the supplementary material.

Theorem 5.1 (Interior point estimates)**:** *Let $\ell(\boldsymbol{\theta})$ be the log-likelihood of 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*

$$\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{L}}) = \arg\max_{\boldsymbol{\theta}\in\Theta}\{\ell(\boldsymbol{\theta}) + P^{FE}(\boldsymbol{\beta}) + P^{MVRE}(\boldsymbol{L})\}\,, \tag{11}$$

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

$$P^{MVRE}(\boldsymbol{L}) = \sum_{i=1}^{N_q} P^{RE}(\log(l_{ii})) + \sum_{i<j}^{q} P^{RE}(l_{ij}), \tag{12}$$

$$P^{RE}(x) \propto \begin{cases} -\frac{1}{2}\{x\}^2, & \text{if } |x| \leq 1 \\ -|x| + \frac{1}{2}, & \text{otherwise} \end{cases}, \tag{13}$$

and

$$P^{FE}(\boldsymbol{\beta}) \propto \log\det(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})\,. \tag{14}$$

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 $\ell(\boldsymbol{\theta}) > -\infty$.

6 Example: conditional inference data

To demonstrate the performance of the MSPAL on a Bernoulli-response GLMM with multivariate random effects structure, we consider a subset of the data analysed by Singmann et al. (2016). As discussed on CrossValidated (`https://stats.stackexchange.com/questions/38493`), this data set exhibits both infinite fixed effects estimates as well as degenerate variance components estimates when a Bernoulli-response GLMM is fitted by MAL.

The data set, originally collected as a control condition of experiment 3)b) in Singmann et al. (2016) and therein analysed in a different context, comes from an experiment in which participants worked on a probabilistic conditional inference task. Participants were presented with the conditional inferences modus ponens (MP), modus tollens (MT), affirmation of the consequent (AC), and denial of the antecedent (DA), for four conditional rules with varying degrees of counterexamples (alternatives, disablers) that are listed below.

1. If a predator is hungry, then it will search for prey. (few disablers, few alternatives)
2. If a person drinks a lot of coke, then the person will gain weight. (many disablers, many alternatives)
3. If a girl has sexual intercourse with her partner, then she will get pregnant. (many disablers, few alternatives)
4. If a balloon is pricked with a needle, then it will quickly loose air. (few disablers, many alternatives)

For each conditional rule and inference, participants were asked to estimate the probability that the conclusion follows from the conditional rule given the minor premise. For example, if MP is "*If p then q. p.*", participants were asked "*If p then q. p. How likely is q?*". Additionally, participants were asked to estimate the probability of the premises themselves. The response variable of this dataset is then a binary response indicating whether, given a certain conditional rule and inference, the participants' probabilistic inference is p-valid; that is, whether their estimate of uncertainty about the conclusion does not exceed the estimated uncertainty of the premises (p-valid inferences are recorded as zero, p-invalid inferences as one). Covariates are the categorical variable counterexamples ("many", "few"), that indicates the degree of available counterexamples to a conditional rule, type ("affirmative","denial") which describes the type of inference (MP and AC are affirmative, MT and DA are denial), and p-validity ("valid","invalid"), indicating whether an inference is p-valid per se (MP and MP are p-valid, while AC and DA are not). For each of the 29 participants, there exist 16 observations corresponding to all possible combinations of inference and conditional rule, giving a total of 464 data points, which are grouped along individuals by the clustering variable code. We can employ a Bernoulli-response GLMM to investigate the probabilistic validity of conditional inference given the type of inference and conditional rule as captured by the covariates and all possible interactions thereof. We introduce a random intercept and random slope for the variable counterexamples to account for response heterogeneity between participants. Hence the model we are considering is given by

$$Y_{ij} \mid \boldsymbol{u}_i \sim \mathrm{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i \tag{15}$$

$$\boldsymbol{u}_i \sim \mathrm{N}(\mathbf{0}_2, \boldsymbol{\Sigma}) \quad (i = 1, \ldots, 29; j = 1, \ldots, 16)\,, \tag{16}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_8)$ are the fixed effects pertaining to the model matrix of the R model formula `response ~ type * p.validity * counterexamples + (1+counterexamples|code)`. As (adaptive) Gauss-Hermite quadrature becomes computationally challenging and not available for `glmer` and consequently `bglmer` for multivariate random effect structures, we approximate the likelihood of model (15) about the parameters $\boldsymbol{\beta}$, $\boldsymbol{L}$ using Laplace's method (see for example Pinheiro and Bates (1995)). We estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{L}$ by MAL using the optimization routines "CG" ("MAL(CG)") and "BFGS" ("MAL(BFGS)") of the `optimx` R package (Nash and Varadhan, 2011), `bglmer` from the `blme` R package Chung et al. (2013) using independent normal ("bglmer(n)") and t ("bglmer(t)") priors for the fixed effects and the default wishart prior for the multivariate variance components. We also estimate the parameters using the proposed MSPAL estimator with the fixed and random effects penalties of Sections 5.1 - 5.2. The estimates are given in Table 3, where we denote the entries of $\boldsymbol{L}$ by l_{ij}, for $i, j = 1, 2$.

Table 3: Estimates from the conditional inference dataset of Singmann et al. (2016) using MAL, `bglmer` and MSPAL

	MAL(BFGS)	MAL(CG)	bglmer(t)	bglmer(n)	MSPAL
β_0	16.25	7.73	13.22	5.45	6.22
	(2.57)	(4.00)	(1.63)	(8.15)	(2.89)
β_2	4.23	3.33	1.86	0.97	0.00
	(1.19)	(14.44)	(3.01)	(2.98)	(4.08)
β_3	-6.69	-2.08	-0.09	-0.13	-2.17
	(1.77)	(2.98)	(1.77)	(2.43)	(2.98)
β_4	-14.40	-5.96	-11.04	-2.88	-4.37
	(2.58)	(4.03)	(1.90)	(8.99)	(2.91)
β_5	3.17	0.85	0.47	0.34	2.17
	(1.36)	(16.40)	(4.54)	(4.32)	(5.02)
β_6	-4.23	-3.20	-1.98	-1.03	0.00
	(1.19)	(14.45)	(3.05)	(3.04)	(4.11)
β_7	8.19	3.81	1.44	1.39	3.64
	(1.83)	(3.11)	(1.94)	(2.56)	(3.09)
β_8	-3.90	-1.86	-1.00	-0.80	-2.87
	(1.91)	(16.43)	(4.66)	(4.44)	(5.12)
$\log l_{11}$	2.02	0.81	4.52	4.52	-0.63
	(0.36)	(1.14)	(0.01)	(0.01)	(2.48)
l_{21}	-7.70	-2.43	-91.89	-92.97	-0.60
	(2.45)	(2.58)	(0.25)	(0.45)	(1.69)
$\log l_{22}$	-5.16	-2.94	-0.27	-0.58	-1.21
	(82.47)	(8.77)	(0.53)	(0.84)	(1.30)

As in the Culcita example of Section 3, we encounter fixed effects estimates that are extreme on the logistic scale for both MAL(BFGS), MAL(CG) and bglmer(t). We further note that the strongly negative estimates for l_{22} in conjunction with the inflated asymptotic standard errors of the MAL(BFGS) estimates are highly indicative of parameter estimates on the boundary of the parameter space, meaning that l_{22} is essentially estimated as zero. The degeneracy of the variance components estimates is even more striking for the estimates using `bglmer`, which give estimates of l_{11}, l_{21} greater than 90 in absolute value, which corresponds to estimated variance components greater than 8000 in absolute value. This underlines that, as with the gamma prior penalty for univariate random effects, the wishart prior penalty, while effective in preventing variance components being estimated as zero, cannot guard against infinite estimates for the variance components. We finally note that for the MSPAL, all parameter estimates as well as their estimated standard errors appear to be finite. Further, while the variance components penalty guards against estimates that are effectively zero, the penalty induced shrinkage towards zero is not as strong as with the whishart prior penalty of the `bglmer` function. To further investigate the frequentist properties of the estimators on this dataset, we repeat the simulation design of the Culcita data example from Section 3 for the conditional inference data where we set the MSPAL estimate of Table 3 as the ground truth. We point out the extremely low percentage of `bglmer` estimates without estimation issues that were used in the summary of Figure 2. While for MSPAL, over 99% of estimates were used in the calculation of the summary statistics of Figure 2, less than 6% were used for the `bglmer` methods. We note that the MSPAL, which is the only estimation method that is guaranteed to give nondegenerate variance components estimates, outperforms MAL and `bglmer`, which incur substantial bias and variance due to their singular and infinite estimates of variance components. Table 3 shows the percentiles of the centered estimates for each estimation method, and underlines that MAL and `blgmer` are unable to guard against degenerate variance components estimates.

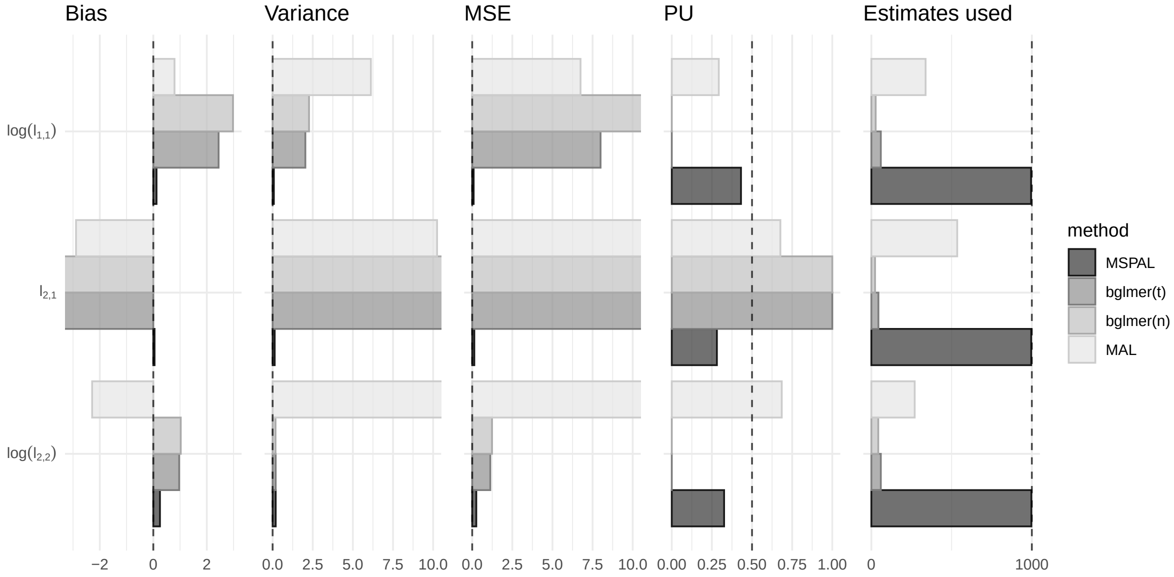

Figure 2: Performance metrics for variance components estimates of MAL,MSPAL and `bglmer` from simulating a Bernoulli-response GLMM from the conditional inference data at the MSPAL

Table 4: Percentiles of centered variance components estimates from simulating a Bernoulli-response GLMM from the conditional inference data at the MSPAL

		Percentiles						
		5%	10%	25%	50%	75%	90%	95%
MSPAL	$\log l_{1,1}$	-0.06	-0.05	-0.02	0.01	0.13	0.46	0.68
	$\log l_{2,2}$	-0.44	-0.32	-0.10	0.25	0.55	0.84	0.98
	$l_{2,1}$	-0.71	-0.26	-0.02	0.12	0.21	0.34	0.42
bglmer(t)	$\log l_{1,1}$	1.03	1.09	1.30	1.79	3.89	4.64	4.95
	$\log l_{2,2}$	0.43	0.53	0.65	0.97	1.11	1.40	1.55
	$l_{2,1}$	-41.35	-34.26	-6.34	-2.67	-1.20	-0.79	-0.69
bglmer(n)	$\log l_{1,1}$	1.24	1.36	1.79	2.03	4.51	4.91	5.12
	$\log l_{2,2}$	0.55	0.59	0.81	1.01	1.17	1.38	1.49
	$l_{2,1}$	-39.80	-32.07	-25.08	-3.06	-2.59	-2.47	-2.09
MAL	$\log l_{1,1}$	-2.48	-1.80	-0.40	1.24	2.36	2.61	2.66
	$\log l_{2,2}$	-7.73	-4.51	-3.40	-1.71	0.28	0.73	1.01
	$l_{2,1}$	-7.49	-7.19	-5.95	-2.45	0.52	0.80	1.06

7 Discussion

This paper proposed the MSPAL estimator for stable parameter estimation in Bernoulli-response GLMMs. We showed that using a scaled version of Jeffreys prior as fixed effects penalty and the negative Huber loss function as a variance components penalty gives nondegenerate estimates whose finite sample properties are superior to the penalized estimator proposed by Chung et al. (2013). While particularly relevant for Bernoulli-response GLMMs, the concept of MSPAL is far more general and we expect it to be useful in other settings, such as GLMMs with Binomial or Poisson responses, for which degenerate M(A)L estimates are known to occur. We leave deriving a unified set of conditions and error rates that satisfy the regularity assumptions that we imposed to derive asymptotic properties of the MSPAL to future research.

References

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.

Bolker, B. (2018). Glmm worked examples, digression: complete separation. GitHub. URL: https://bbolker.github.io/mixedmodels-misc/ecostats_chap.html#digression-complete-separation.

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, and V. J. Sosa (Eds.), *Ecological Statistics*, pp. 309–333. Oxford University Press.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution 24*(3), 127–135.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*(421), 9–25.

Browne, W. J. and D. Draper (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis 1*(3), 473–514.

Charles E. McCulloch, Shayle R. Searle, J. M. N. (2008). *Generalized, linear, and mixed models*, Volume 2. John Wiley & Sons.

Chung, Y., A. Gelman, S. Rabe-Hesketh, J. Liu, and V. Dorie (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics 40*(2), 136–157.

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.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*(4), 1360 – 1383.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*(16), 2409–2419.

Jiang, J. (2017). *Asymptotic analysis of mixed effects models: theory, applications, and open problems.* Chapman and Hall/CRC.

Jin, S. and B. Andersson (2020). A note on the accuracy of adaptive gauss–hermite quadrature. *Biometrika 107*(3), 737–744.

Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models.* Ph. D. thesis, University of Oxford.

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

Kosmidis, I. and D. Schumacher (2021). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates.* R package version 0.2.

Liu, Q. and D. A. Pierce (1994). A note on Gauss-Hermite quadrature. *Biometrika 81*(3), 624–629.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

McKeon, C. S., A. C. Stier, S. E. McIlroy, and B. M. Bolker (2012). Multiple defender effects: synergistic coral defense by mutualist crustaceans. *Oecologia 169*(4), 1095–1103.

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.

Ogden, H. (2019). *glmmsr: Fit a Generalized Linear Mixed Model.* R package version 0.2.3.

Ogden, H. (2021). On the error in laplace approximations of high-dimensional integrals. *Stat 10*(1), e380.

Pasch, B., B. M. Bolker, and S. M. Phelps (2013). Interspecific dominance via vocal interactions mediates altitudinal zonation in neotropical singing mice. *The American Naturalist 182*(5), E161–E173.

Pinheiro, J. C. and D. M. Bates (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics 4*(1), 12–35.

Pinheiro, J. C. and E. C. Chao (2006). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics 15*(1), 58–81.

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

Raudenbush, S. W., M.-L. Yang, and M. Yosef (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics 9*(1), 141–157.

Rigollet, P. (2015). *18.S997 High-Dimensional Statistics.* https://ocw.mit.edu: Massachusetts Institute of Technology: MIT OpenCourseWare.

Rodriguez, G. and N. Goldman (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 158*(1), 73–89.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Singmann, H., K. C. Klauer, and S. Beller (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology 88*, 61–87.

Stringer, A. and B. Bilodeau (2022). Fitting generalized linear mixed models using adaptive quadrature.

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

Wolfinger, R. and M. O'connell (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation 48*(3-4), 233–243.

Zehna, P. W. (1966). Invariance of Maximum Likelihood Estimators. *The Annals of Mathematical Statistics 37*(3), 744–744.

Zhao, Y., J. Staudenmayer, B. A. Coull, and M. P. Wand (2006). General Design Bayesian Generalized Linear Mixed Models. *Statistical Science 21*(1), 35 – 51.

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 $\sup_{\boldsymbol{\theta}\in\Theta} \left|\left|\left|r_n^{-1}J(\boldsymbol{\theta}) - I(\boldsymbol{\theta})\right|\right|\right| \xrightarrow{p} 0$ 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 $\mu_j(\boldsymbol{\beta}) = \operatorname{logit}^{-1}(\boldsymbol{x}_j^\top\boldsymbol{\beta}) \in (0,1)$.

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 $[\widetilde{\widetilde{\boldsymbol{W}}}_{ij}]_{kk} = \left[(1-2\mu_k)(1-3\mu_k)-\mu_k\right]x_{ki}x_{kj}$, 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

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

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

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

Hence

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

So that it follows that

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

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

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

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

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

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

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

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

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 $i = 1, \dots, 5$ and within cluster observations $j = 1, \dots, n$, $n \in \{50, 100, 200\}$, 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 $X_{i1} = 1, X_{i2} \sim \mathrm{N}(0,1), X_{i3} \sim \text{ Ber}\left(\frac{1}{2}\right), X_{i4} \sim \text{ Ber}\left(\frac{1}{4}\right)$, and $X_{i5} \sim \text{ exp}(1)$. 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

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

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.

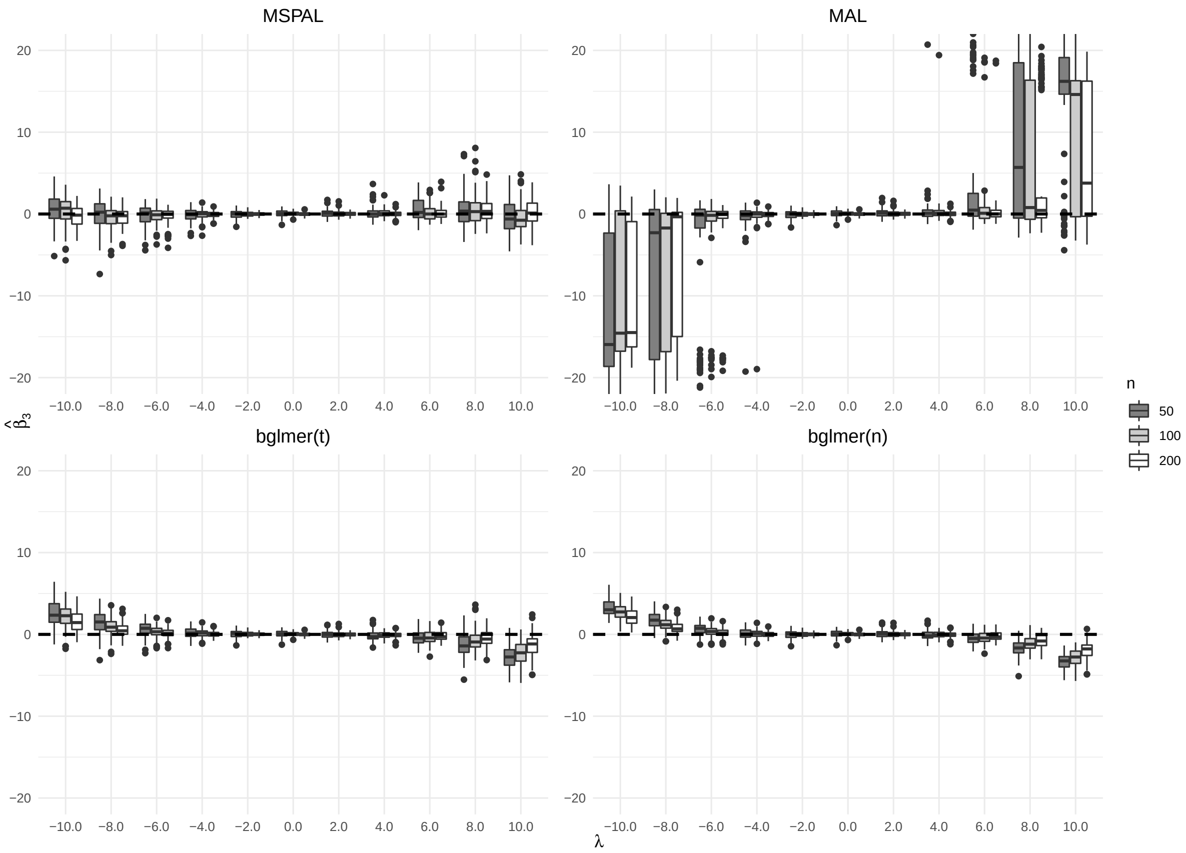

Figure S1: Centered estimation output of $\hat{\beta}_3 - \beta_3$ from Simulation 1

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 $i = 1, \dots, 5$ and within cluster observations $j = 1, \dots, n$, $n \in \{50, 100, 200\}$, 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 $X_{i1} = 1$, $X_{i2} \sim \mathrm{N}(0, 1), X_{i3} \sim \text{Ber}\left(\frac{1}{2}\right), X_{i4} \sim \text{Ber}\left(\frac{1}{4}\right)$, 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

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

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.

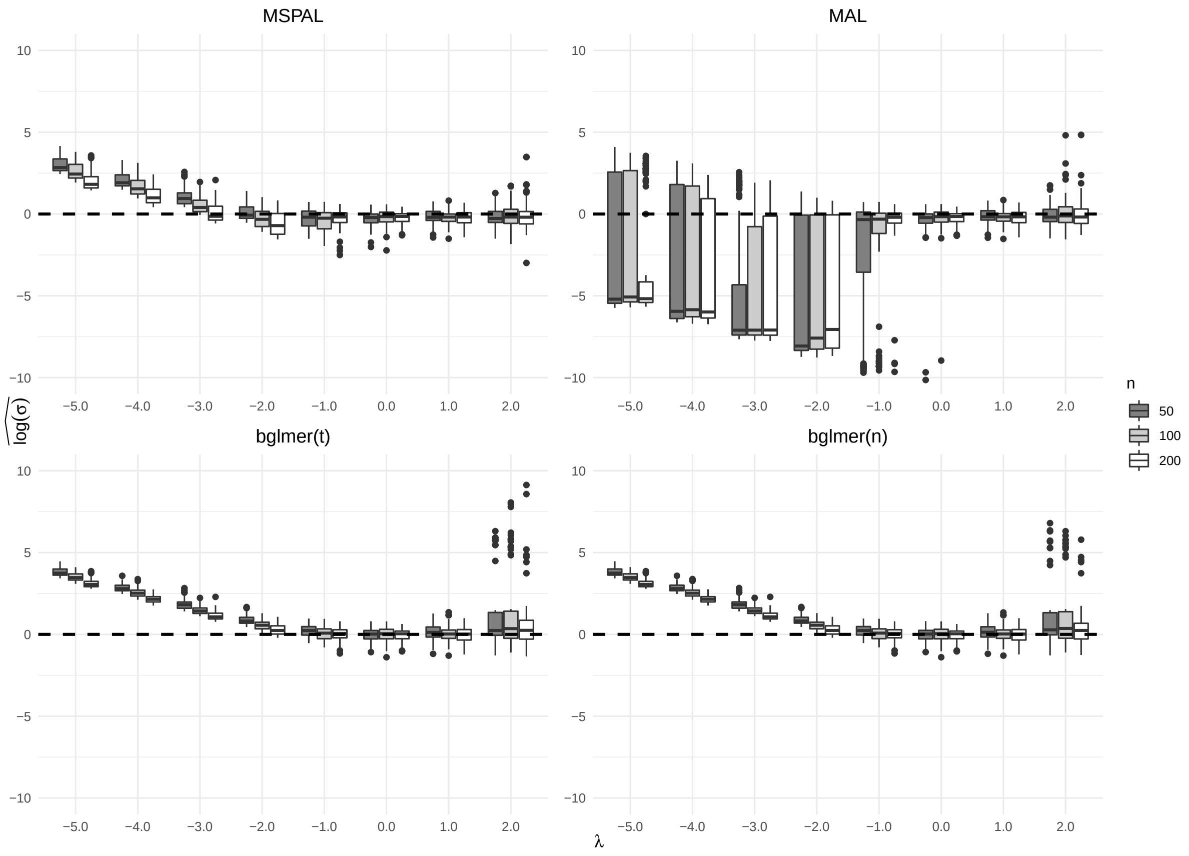

Figure S2: Centered estimation output of $\widehat{\log \sigma} - \log(\sigma)$ from Simulation 2

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.