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 8, 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 parameters 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 guaranteed to be 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 parameter 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

sec:intro

Generalized Linear Mixed Models (GLMMs; McCulloch and Searle 2004, 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 therefore maximize an attainable 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). 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 parameters 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 **#PS: not familiar with term 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; Shen and Gao, 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 employed for establishing consistency and asymptotic normality 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) **#IK: add citations**. Maximum softly-penalized likelihood is compared to prominent competing approaches through two real-data examples,

and simulation studies that illustrate its superior finite-sample performance. **#PS: we only compare to MAL and bglmer, does this warrant this formulation?**

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, that are aimed at provoking particular degenerate MAL estimates, are given in the supplementary material.

2 Bernoulli-response generalized linear mixed models

sec:bern_GLMMs

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,

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

eq:bern_cluster

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

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 effect parameters $\boldsymbol{\beta} \in \Re^p$, 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. **#PS: The next sentence needs discussion, I do not understand this formulation. Shouldn't it be: The other often encountered formulation of GLMMs in the literature (McCulloch and Searle, 2004, Chapter 7.4) absorbs the clustering into the variance components structure of $\boldsymbol{\Sigma}$ and is therefore a clustered GLMM with a single cluster.**The other version **#IK: give a reference** that often appears in the literature is then simply a clustered GLMM with one observation per 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

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

eq:bern_likl

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).

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`).

tab:culcita

	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

3 Motivating example

sec:culcita_dat

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.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

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

eq:logistic_normal

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

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

Table 2: **#IK: write this**

tab:culcita_inf

	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)

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.

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 vary-

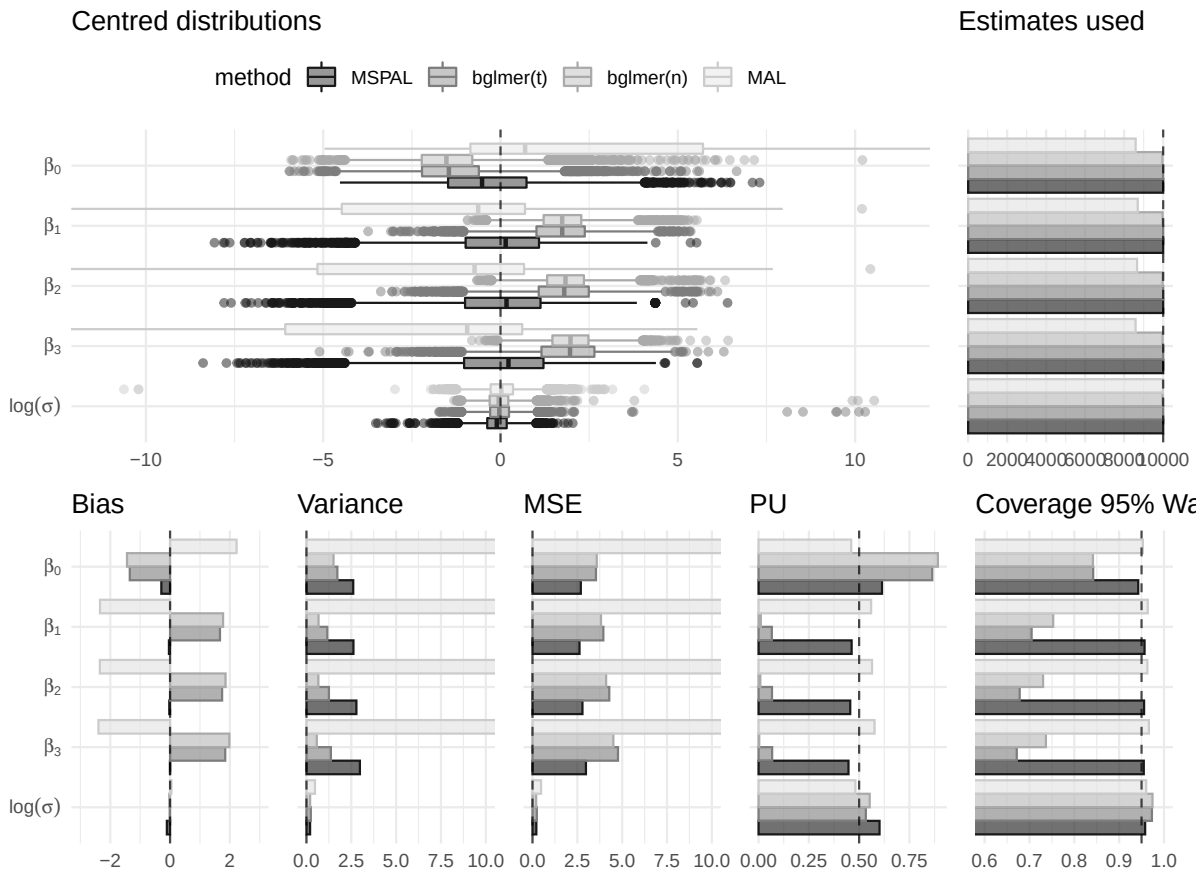

Figure 1: **#IK: write this** fig:culcita_simu0

ing 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 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 MSAPL 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. **#PS: how do wee see the asymptotics(optimality) from figure 1?**

4 Penalized likelihoods

sec:softpen

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

sec:interior

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 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

sec:ass+res

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 $\tilde{S}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$.

thm:soft_pen_cons

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

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 $\boldsymbol{\theta}_0$

thm:asymp_norm_soft_pen

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.

thm:hypo

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

sec:soft_pen

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 $\|\nabla P_u(\boldsymbol{\theta})\|$, $\left\|\left|\nabla\nabla^{\top} P_u(\boldsymbol{\theta})\right|\right\|$ 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 parameters 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

sec:glmm_penalties

5.1 Fixed effects penalty

sec:glmm_fe_pen

The unscaled fixed effects parameter 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 $\boldsymbol{\beta} = \boldsymbol{0}_p$ 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

$$\left|\frac{\partial}{\partial\beta_i}\log\det(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X})\right| \leq p\max_{1\leq j\leq n}|x_{ji}| \qquad \text{eq:jeffrey\_deriv\_bound} \quad (6)$$

$$\left|\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log\det(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X})\right| \leq 2p\max_{1\leq k\leq n}|x_{ki}|\max_{1\leq k\leq n}|x_{kj}| \qquad \text{eq:jeffrey\_deriv\_bound2} \quad (7)$$

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

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

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 $\max_{i,j}|x_{ji}| = \mathcal{O}_p(\sqrt{2\sigma^2\log(2np)})$ (see for example Rigollet (2015, Theorem 1.14)).

5.2 Variance components penalty

sec:glmm_re_pen

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

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

Following the discussion of Section 4.5, the variance components penalty of (9) must satisfy $\lim_{\sigma\to 0} P_u^{\mathrm{RE}}(\sigma) = -\infty$ 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 $\lim_{\sigma\to 0} P_u^{\mathrm{RE}}(x) = -\infty$. 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

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

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.

thm:int_point

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*

eq:pen_max

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

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

sec:ci

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

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

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

eq:cond_inf_model

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. (2015) 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

tab:cond_inf

	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.

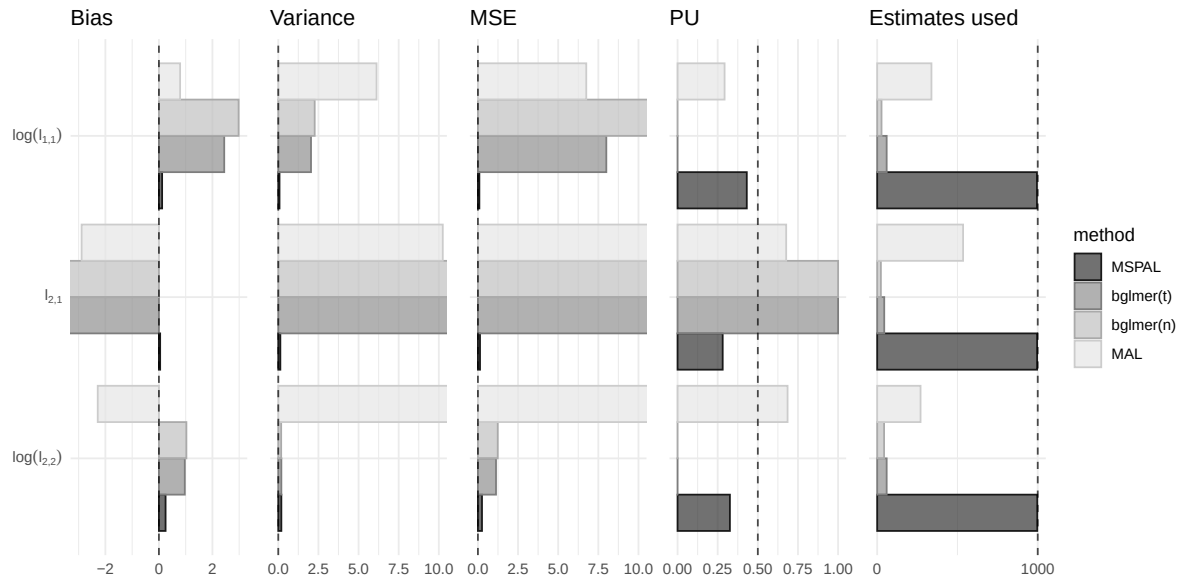

Figure 2: Performance metrics for variance components estimates of MAL,MSPAL and `bglmer` from simulating a Bernoulli-response GLMM from the conditional inference data at the MSPAL

fig:cond_inf_simul

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

tab:sim2

		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

sec:sum

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.

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.). Number 37. Boca Raton: Chapman & Hall/CRC.

McCulloch, C. E. and S. R. Searle (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

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).

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.

Shen, J. and S. Gao (2008). A solution to separation and multicollinearity in multiple logistic regression. *Journal of data science: JDS 6*(4), 515.

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. *arXiv preprint arXiv:2202.07864*.

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*, 35–51.