Maximum softly-penalized likelihood for Bernoulli-response GLMMs

Philipp Sterzinger, Ioannis Kosmidis

Department of Statistics, University of Warwick, Coventry, UK
philipp.sterzinger@warwick.ac.uk

WARWICK
THE UNIVERSITY OF WARWICK

We provide a computationally stable estimation method for Bernoulli-response GLMMs that returns interior point estimates and has optimal asymptotic properties

Bernoulli-response GLMMs

Model

$$Y_{ij} \mid \boldsymbol{u}_i \sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i$$
$$\boldsymbol{u}_i \sim \text{N}(\boldsymbol{0}_q, \boldsymbol{\Sigma}) \quad (i = 1, \dots, k; j = 1, \dots, n_i)$$

where $\mu_{ij} = P(Y_{ij} = 1 \mid \boldsymbol{u}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, $g : (0,1) \to \Re$ is a known link function, e.g. $g(x) = \log(x/1-x)$, and $\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}$ are vectors of fixed and random effect covariates.

Marginal likelihood about $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$

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

Maximum approximate likelihood (MAL) estimator

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\ell(\boldsymbol{\theta})$ is the logarithm of an approximation to (1).

Maximum softly penalized likelihood

Maximum softly penalized approximate likelihood (MSPAL) estimator

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

where $P(\boldsymbol{\theta})$ is a penalty that satisfies $\lim_{r \to \infty} P(\boldsymbol{\theta}(r)) = -\infty$, for any sequence of parameters $\boldsymbol{\theta}(r)$ such that $\lim_{r \to \infty} \boldsymbol{\theta}(r)$ lies on the boundary.

Then, if there is a point $\boldsymbol{\theta}$ such that $\ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) > -\infty$, $\tilde{\boldsymbol{\theta}}$ is in the interior of the parameter space.

"Soft" penalty

To preserve the MAL asymptotics, we adapt the work of [3] and control $\|\nabla P(\boldsymbol{\theta})\|$ in terms of the rate of information accumulation r_n, for which the observed information $J(\boldsymbol{\theta}) = -\nabla\nabla^\top \ell(\boldsymbol{\theta})$ converges to a nonrandom $\mathcal{O}(1)$ matrix $I(\boldsymbol{\theta})$, i.e. $r_n^{-1} J(\boldsymbol{\theta}) \xrightarrow{p} I(\boldsymbol{\theta})$.

Consistency

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

Theorem 1

Let $\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla P(\boldsymbol{\theta})\| = o_p(r_n)$, and assume that A0-A3 hold. Then $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Asymptotic Normality

A4 Both $\ell(\boldsymbol{\theta}), \tilde{\ell}(\boldsymbol{\theta})$ are three times differentiable
A5 $\sup_{\boldsymbol{\theta} \in \Theta} \left\| \left| r_n^{-1} J(\boldsymbol{\theta}) - I(\boldsymbol{\theta}) \right| \right\| \xrightarrow{p} 0$ and $I(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ in a neighbourhood around $\boldsymbol{\theta}_0$
A6 $r_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \text{N}(0, I(\boldsymbol{\theta}_0)^{-1})$
A7 $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$

Theorem 2

Let $\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla P(\boldsymbol{\theta})\| = o_p(r_n^{1/2})$ and assume that conditions A3-A7 hold. Then

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

Soft penalties for Bernoulli-response GLMMs

We propose to penalize the approximate log-likelihood $\ell(\boldsymbol{\theta})$ by a composite penalty that penalizes the fixed effects $\boldsymbol{\beta}$ and the Cholesky factor of the variance components, say $\boldsymbol{L}$, separately, i.e.

$$P(\boldsymbol{\theta}) = P^{FE}(\boldsymbol{\beta}) + P^{\text{VC}}(\boldsymbol{L})$$

Fixed effects penalty

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

where $\boldsymbol{X}$ is the $n \times p$ matrix of all fixed effect covariates, $\boldsymbol{W}$ is a diagonal matrix with entries $\boldsymbol{W}_{ii} = \mu_i(\boldsymbol{\beta})(1 - \mu_i(\boldsymbol{\beta}))$, $\log(\mu_i(\boldsymbol{\beta})/(1-\mu_i(\boldsymbol{\beta})) = \boldsymbol{x}_i^\top \beta$.

[4] show that $\lim_{r \to \infty} P^{FE}(\boldsymbol{\beta}(r)) = -\infty$ for any sequence of fixed effects such that $\lim_{r \to \infty} \boldsymbol{\beta}(r)$ has infinite components, which guarantees finite MSPAL fixed effects estimates.

We show that $\sup_{\boldsymbol{\beta} \in \Re^p} \|\nabla P^{FE}(\boldsymbol{\beta})\| = \mathcal{O}(1)$, whenever $\max_{i,j} |x_{ij}| = \mathcal{O}(n^{1/2})$ in line with Theorems 1 & 2.

Variance components penalty

$$P^{\text{VC}}(\boldsymbol{L}) = \sum_{i=1}^{q} P^{\text{V}}(\log(l_{ii})) + \sum_{i<j}^{q} P^{\text{VC}}(l_{ij}), \quad P^{\text{V}}(x) = \sqrt{p/n} \begin{cases} -\frac{1}{2}\{x\}^2 & \text{if } |x| \leq 1, \\ -|x| + \frac{1}{2} & \text{otherwise} \end{cases}$$

where $\boldsymbol{L}$ is the lower-triangular Cholesky factor of the variance components, so that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. This penalty gives estimated variance components matrix $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^\top$ that are finite, nonsingular and without perfect estimated correlation, i.e. for all $i \neq j$, $\left| \frac{\tilde{\boldsymbol{\Sigma}}_{ij}}{\sqrt{\tilde{\boldsymbol{\Sigma}}_{ii}\tilde{\boldsymbol{\Sigma}}_{jj}}} \right| < 1$.

Motivating Example

Data

Table 1: Culcita data from the worked examples of [1]

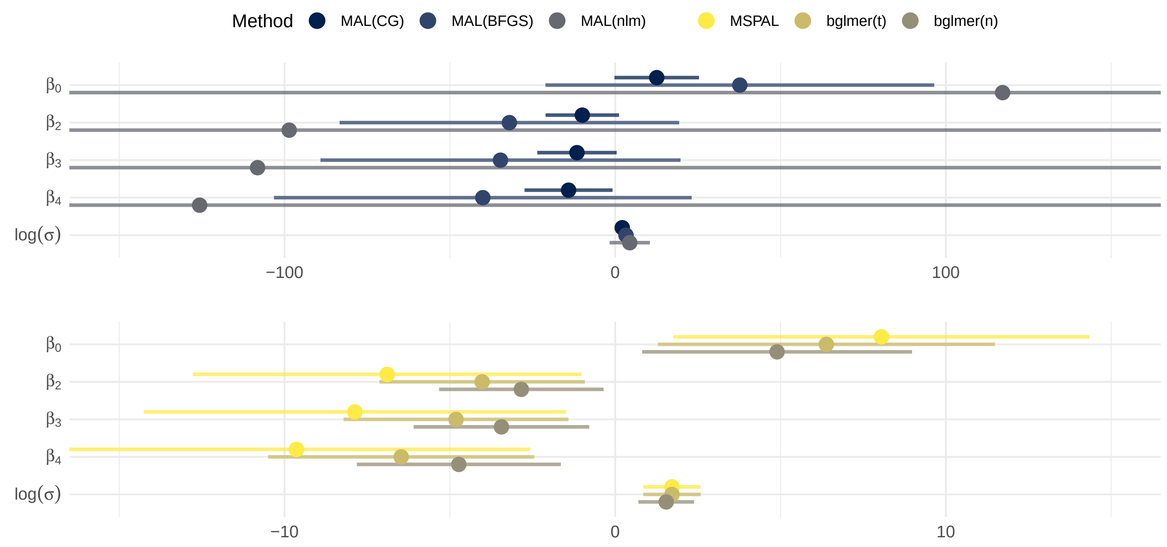
	Block									
Treatment	1	2	3	4	5	6	7	8	9	10
none	0,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,0
crabs	0,0	0,0	0,0	0,0	1,1	1,1	1,1	1,1	1,1	1,1
shrimp	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1	1,1
both	0,0	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1

The complete randomized block design (four treatments, ten temporal blocks, two replications) data records coral-eating sea stars (Culcita) attacking coral harbouring different protective symbionts (crabs, shrimp).

Model

Upon removal of the atypical observation in block 10 without predation and symbionts, we associate predation to treatment effects using a Bernoulli-response GLMM with logistic link and a random intercept per block, so that $\eta_{ij} = \beta_0 + u_i + \beta_j$, $u_i \sim \text{N}(0, \sigma^2)$ for $i = 1, \dots, 10, j = 1, \dots, 4$. β_1 is set to zero, making treatment "none" the reference category.

Infinite & disparate estimates



We estimate $\boldsymbol{\beta}, \log\sigma$ by MAL, with a 200-point Gauss-Hermite quadrature approximation to the log-likelihood and three different optimization routines from the R `optimx` package, and our proposed MSPAL estimator and `bglmer` [2], with a 100-point adaptive Gauss-Hermite quadrature approximation to the log-likelihood.

The MAL estimates and asymptotic 95% CIs about the model parameters are highly dissimilar and effectively infinite but, due to different stopping criteria of the optimizers, may not be perceived as such. MSPAL and `bglmer` estimates are finite, with `bglmer` exhibiting stronger shrinkage.

Simulation study

Simulation results for the MAL, MSPAL and the `bglmer` routine of [2] from 10000 independent samples of responses at the MAL estimates when all data points are used.

References

[1] B. M. Bolker, "Linear and generalized linear mixed models," in *Ecological Statistics*, Oxford University Press, 2015, pp. 309–333. DOI: 10.1093/acprof:oso/9780199672547.003.0014.
[2] Y. Chung, S. Rabe-Hesketh, *et al.*, "A nondegenerate penalized likelihood estimator for variance parameters in multilevel models," *Psychometrika*, vol. 78, no. 4, pp. 685–709, 2013.
[3] H. Ogden, "On asymptotic validity of naive inference with an approximate likelihood," *Biometrika*, vol. 104, no. 1, pp. 153–164, 2017.
[4] I. Kosmidis and D. Firth, "Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models," *Biometrika*, vol. 108, no. 1, pp. 71–82, 2021.