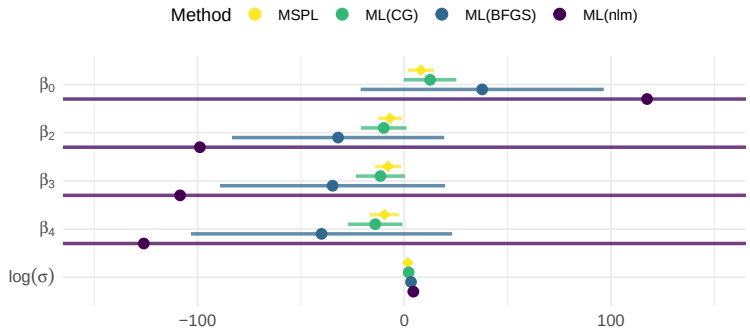
WARWICK
THE UNIVERSITY OF WARWICK

Maximum softly-penalized likelihood
for mixed effects logistic regression

Philipp Sterzinger
philipp.sterzinger@warwick.ac.uk

Maximum softly-penalized likelihood for mixed effects logistic regression

We provide a penalized likelihood estimation method for mixed effects logistic regression that gives estimates in the interior of the parameter space with optimal asymptotic properties



Mixed effects logistic regression
Model

- Response vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top \in \{0,1\}^{n_i}$, and sequence of covariates $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k$, with $\boldsymbol{V}_i \in \Re^{n_i \times s}$
- $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are realizations of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, whose entries $Y_{i1}, \ldots, Y_{in_i}$, are independent Bernoulli random variables conditionally on a vector of random effects $\boldsymbol{u}_i$
- $\boldsymbol{u}_i$ are independent realizations of a multivariate normal distribution

$$
\begin{aligned}
Y_{ij} \mid \boldsymbol{u}_i &\sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i \\
\boldsymbol{U}_i &\sim \mathrm{N}(\boldsymbol{0}_q, \boldsymbol{\Sigma}) \quad (i = 1, \ldots, k; j = 1, \ldots, n_i)\,,
\end{aligned}
\tag{1}
$$

where $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$ are the jth row of matrices $\boldsymbol{X}_i, \boldsymbol{Z}_i$ constructed from columns of $\boldsymbol{V}_i$, and $\boldsymbol{\beta} \in \Re^p, \boldsymbol{\Sigma} \in \Re^{q \times q}$

Mixed effects logistic regression
Marginal likelihood

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

Marginal likelihood about $\boldsymbol{\beta}, \boldsymbol{\Sigma}$:

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

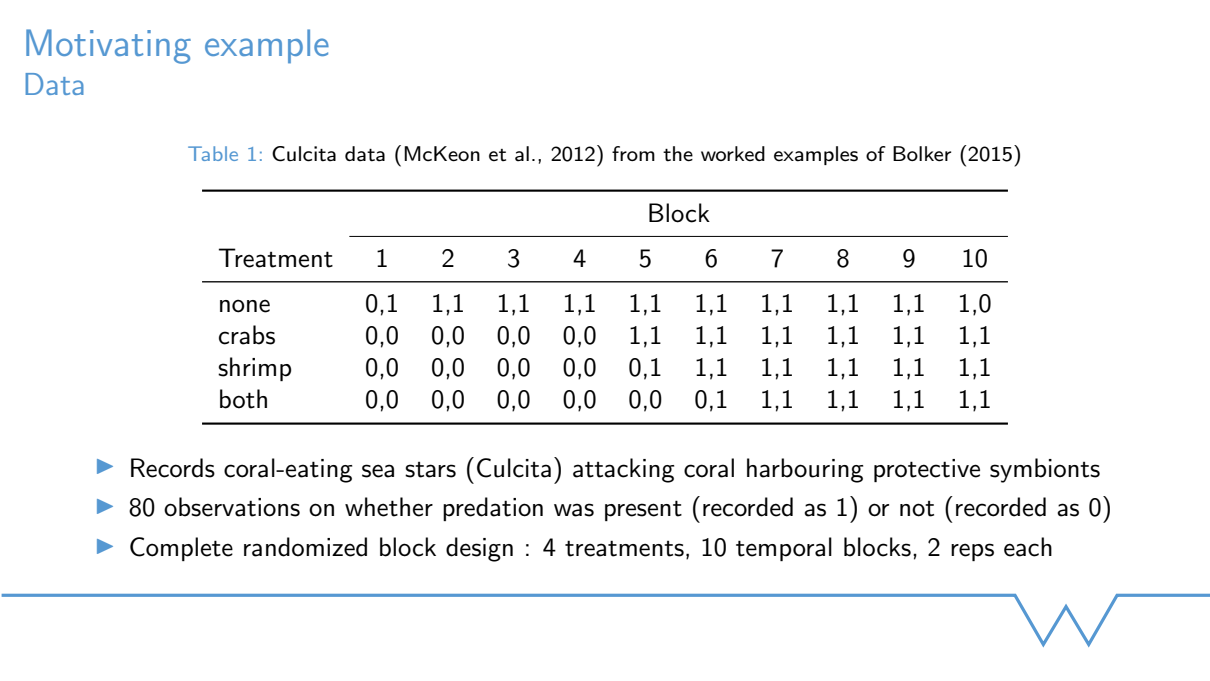
- Maximum likelihood estimator: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}} = \arg\max_{\boldsymbol{\beta}, \boldsymbol{\Sigma}} L(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ (numerical approx. of (2))
- Estimates can lie on the boundary of the parameter space (infinite $\hat{\boldsymbol{\beta}}$, infinite/singular $\hat{\boldsymbol{\Sigma}}$)

Motivating example
Data

Table 1: Culcita data (McKeon et al., 2012) from the worked examples of Bolker (2015)

	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

- Records coral-eating sea stars (Culcita) attacking coral harbouring protective symbionts
- 80 observations on whether predation was present (recorded as 1) or not (recorded as 0)
- Complete randomized block design : 4 treatments, 10 temporal blocks, 2 reps each

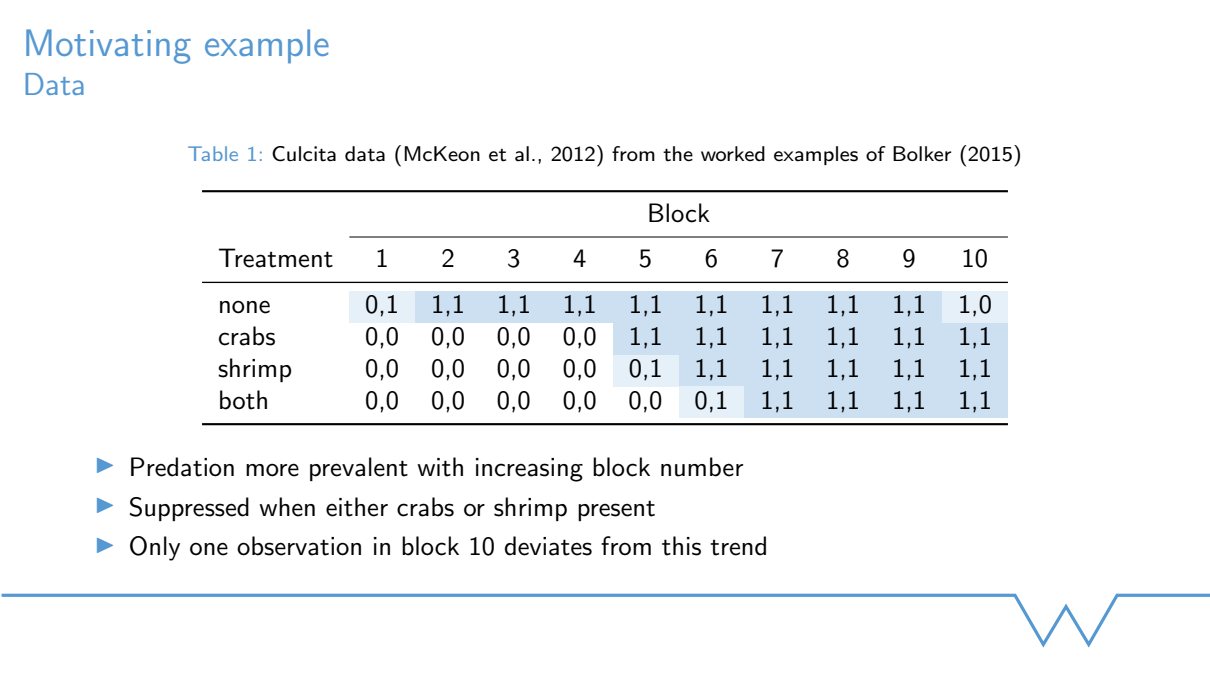


Motivating example
Data

Table 1: Culcita data (McKeon et al., 2012) from the worked examples of Bolker (2015)

	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

- Predation more prevalent with increasing block number
- Suppressed when either crabs or shrimp present
- Only one observation in block 10 deviates from this trend



Motivating example
Model

Associate predation with treatment effects while accounting for heterogeneity between blocks:

$$Y_{ij} \mid u_i \sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = \beta_0 + \beta_{a(j)} + u_i \tag{3}$$
$$u_i \sim \text{N}(0, \sigma^2) \quad (i = 1, \ldots, 10; j = 1, \ldots, 8),$$

where $a(j) = \lceil j/2 \rceil$ and $(Y_{i1}, Y_{i2})^\top, (Y_{i3}, Y_{i4})^\top, (Y_{i5}, Y_{i6})^\top, (Y_{i7}, Y_{i8})^\top$ correspond to the two responses for each of "none", "crabs", "shrimp", and "both"

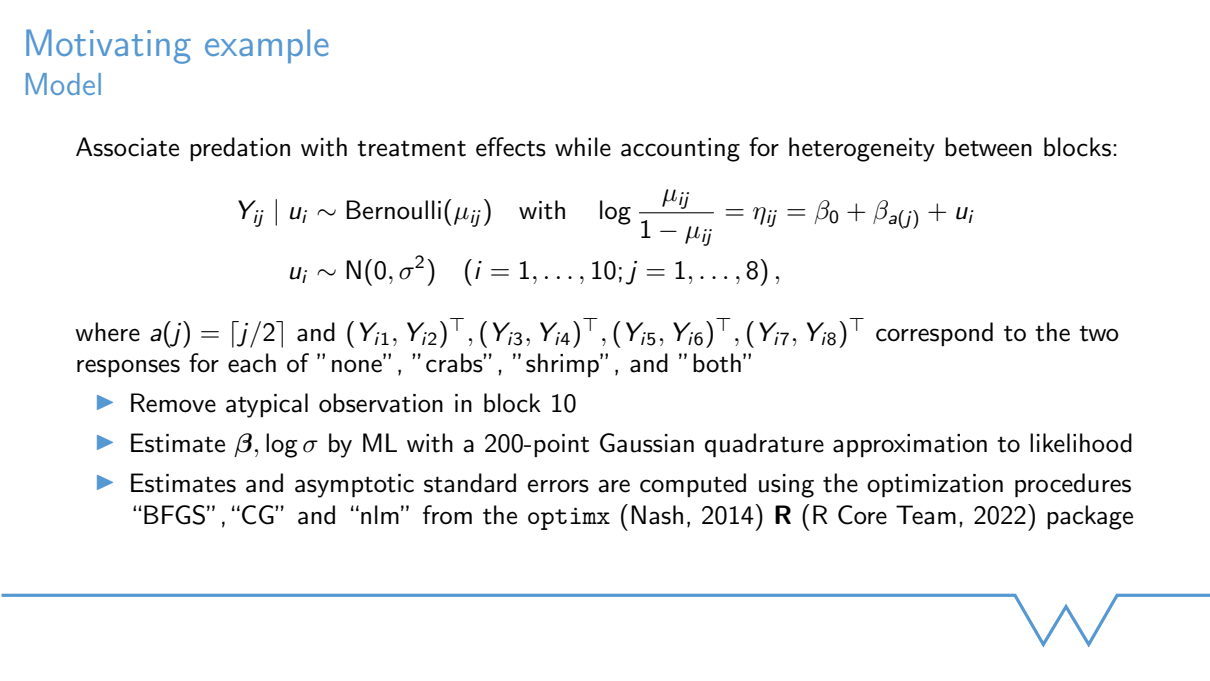
Motivating example
Model

Associate predation with treatment effects while accounting for heterogeneity between blocks:

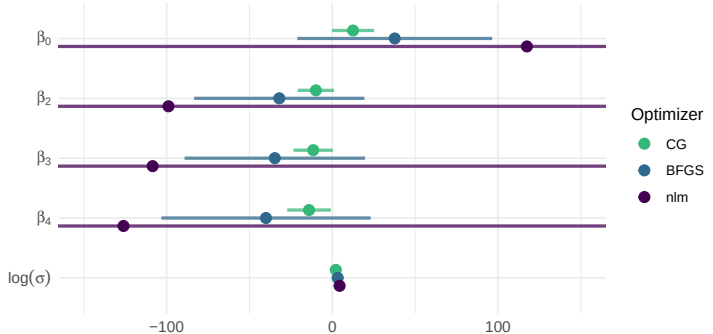
$$Y_{ij} \mid u_i \sim \text{Bernoulli}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = \beta_0 + \beta_{a(j)} + u_i$$
$$u_i \sim \text{N}(0, \sigma^2) \quad (i = 1, \ldots, 10; j = 1, \ldots, 8),$$

where $a(j) = \lceil j/2 \rceil$ and $(Y_{i1}, Y_{i2})^\top, (Y_{i3}, Y_{i4})^\top, (Y_{i5}, Y_{i6})^\top, (Y_{i7}, Y_{i8})^\top$ correspond to the two responses for each of "none", "crabs", "shrimp", and "both"

- Remove atypical observation in block 10
- Estimate $\boldsymbol{\beta}, \log \sigma$ by ML with a 200-point Gaussian quadrature approximation to likelihood
- Estimates and asymptotic standard errors are computed using the optimization procedures "BFGS", "CG" and "nlm" from the `optimx` (Nash, 2014) **R** (R Core Team, 2022) package



Motivating example
Estimates



Estimates are different and extreme on the logistic scale

- Early stoppage of optimization routines after prematurely declaring convergence
- Large standard errors indicate almost flat approximate log-likelihood around the estimates
- The MAL estimates for $\beta_0, \beta_1, \beta_2, \beta_3$ are in reality infinite in absolute value

Maximum softly-penalized likelihood

Setup

Penalize log-likelihood with penalty that diverges to $-\infty$ when we approach the parameter space boundary (cf. Chung et al. (2013))

Maximum softly-penalized likelihood

Setup

Penalize log-likelihood with penalty that diverges to $-\infty$ when we approach the parameter space boundary (cf. Chung et al. (2013))

- Reparametrize $\boldsymbol{\Sigma} = s(\boldsymbol{\psi})$, for

$$\boldsymbol{\psi} = (\log l_{11}, \ldots, \log l_{qq}, l_{21}, \ldots, l_{q1}, l_{32}, \ldots, l_{q2}, \ldots, l_{qq-1})^\top ,$$

  where l_{ij} $(i > j)$ is the (i,j)th element of $\boldsymbol{L}$, and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$
- Model parameters: $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\psi}^\top)^\top$, $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\beta}, s(\boldsymbol{\psi}))$, where $L(\boldsymbol{\beta}, s(\boldsymbol{\psi}))$ is (2)
- MPL estimator: $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \{\ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta})\}$ for some penalty $P(\boldsymbol{\theta})$
- Composite penalty:

$$P(\boldsymbol{\theta}) = c_1 P_{(f)}(\boldsymbol{\beta}) + c_2 P_{(v)}(\boldsymbol{\psi}) ,$$

  where $c_1 > 0$, $c_2 > 0$, and $P_{(f)}(\boldsymbol{\beta})$ and $P_{(v)}(\boldsymbol{\psi})$ are unscaled penalty functions for the fixed effects and variance components

Maximum softly-penalized likelihood

Fixed Effects Penalty

Logarithm of Jeffreys' invariant prior for the corresponding GLM, i.e.

$$P_{(f)}(\boldsymbol{\beta}) = \frac{1}{2} \log \det \left(\sum_{i=1}^{k} \boldsymbol{X}_i^\top \boldsymbol{W}_i \boldsymbol{X}_i \right)$$

- $\boldsymbol{X}_i$ collects the fixed effects covariates of cluster i in model (1),
- $\boldsymbol{W}_i$ is a diagonal matrix with jth element $\mu_{ij}^{(f)}(1 - \mu_{ij}^{(f)})$, $\mu_{ij}^{(f)} = 1/\{1 + \exp(-\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})\}$

Variance components penalty

Composition of negative Huber loss functions on the components of $\boldsymbol{\psi}$

$$P_{(v)}(\boldsymbol{\psi}) = \sum_{i=1}^{q} D(\log l_{ii}) + \sum_{i>j} D(l_{ij}), \quad D(x) = \begin{cases} -\frac{1}{2}x^2, & \text{if } |x| \leq 1 \\ -|x| + \frac{1}{2}, & \text{otherwise} \end{cases}$$

Maximum softly-penalized likelihood

Non-boundary estimates

Denote by $\partial\Theta$ the boundary of Θ and let $\boldsymbol{\theta}(r)$, $r \in \Re$, be a path in the parameter space such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) \in \partial\Theta$

If

- $P(\boldsymbol{\theta})$ is such that $\lim\limits_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$, and
- $\ell(\boldsymbol{\theta})$ is bounded from above and not $-\infty$ for all $\boldsymbol{\theta} \in \Theta$,

then $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\psi}})$ lies in the interior of the parameter space, i.e.

- All components of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Sigma}} = s(\tilde{\boldsymbol{\psi}})$ are finite,
- $\tilde{\boldsymbol{\Sigma}}$ is positive definite, with implied correlations away from -1 and 1

Maximum softly-penalized likelihood

Non-boundary estimates

Denote by $\partial\Theta$ the boundary of Θ and let $\boldsymbol{\theta}(r)$, $r \in \Re$, be a path in the parameter space such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) \in \partial\Theta$

If

- $P(\boldsymbol{\theta})$ is such that $\lim\limits_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$, and
- $\ell(\boldsymbol{\theta})$ is bounded from above and not $-\infty$ for all $\boldsymbol{\theta} \in \Theta$,

then $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\psi}})$ lies in the interior of the parameter space, i.e.

- All components of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Sigma}} = s(\tilde{\boldsymbol{\psi}})$ are finite,
- $\tilde{\boldsymbol{\Sigma}}$ is positive definite, with implied correlations away from -1 and 1

✔ Satisfied by our composite penalty (see Kosmidis and Firth (2021) for fixed effects penalty)

Maximum softly-penalized likelihood

Equivariance under linear transformations of fixed effects

ML estimates are equivariant under transformations of model parameters (Zehna, 1966)

- Particularly useful: Scaled linear transformations $\boldsymbol{\beta}' = \boldsymbol{C}\boldsymbol{\beta}$ of the fixed effects for known, invertible, real matrices $\boldsymbol{C}$
- Obtain ML estimates and standard errors for arbitrary sets of scaled parameter contrasts, when estimates for one of those sets of contrasts are available without re-estimating the model
- For MPL require:

$$P_{(f)}(\boldsymbol{C}\boldsymbol{\beta}) = P_{(f)}(\boldsymbol{\beta}) + a\,,$$

where $a \in \Re$ is a scalar that does not depend on $\boldsymbol{\beta}$

Maximum softly-penalized likelihood

Equivariance under linear transformations of fixed effects

ML estimates are equivariant under transformations of model parameters (Zehna, 1966)

- Particularly useful: Scaled linear transformations $\boldsymbol{\beta}' = \boldsymbol{C}\boldsymbol{\beta}$ of the fixed effects for known, invertible, real matrices $\boldsymbol{C}$
- Obtain ML estimates and standard errors for arbitrary sets of scaled parameter contrasts, when estimates for one of those sets of contrasts are available without re-estimating the model
- For MPL require:

$$P_{(f)}(\boldsymbol{C}\boldsymbol{\beta}) = P_{(f)}(\boldsymbol{\beta}) + a\,,$$

where $a \in \Re$ is a scalar that does not depend on $\boldsymbol{\beta}$

✔ For Jeffreys' prior, $P_{(f)}(\boldsymbol{C}\boldsymbol{\beta}) = P_{(f)}(\boldsymbol{\beta}) - \log\det \boldsymbol{C}$, so equivariance holds in contrast to other MPL approaches such as `bglmer`'s default penalties

Maximum softly-penalized likelihood

Consistency and asymptotic normality

To preserve ML asymptotics, scale $P(\boldsymbol{\theta})$ to make penalization asymptotically negligible

Maximum softly-penalized likelihood

Consistency and asymptotic normality

To preserve ML asymptotics, scale $P(\boldsymbol{\theta})$ to make penalization asymptotically negligible

- Choose scaling factors c_1, c_2 to control $\|\nabla P(\boldsymbol{\theta})\|$ in terms of the rate of information accumulation r_n, which is such that

$$r_n^{-1}\left\{-\nabla\nabla^\top \ell(\boldsymbol{\theta})\right\} \xrightarrow{p} I(\boldsymbol{\theta}), \quad I(\boldsymbol{\theta}) = \mathcal{O}(1)$$

- For $c_1 = c_2 = 2\sqrt{p/n}$, $n = \sum_{i=1}^{k} n_i$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla P(\boldsymbol{\theta})\| \leq \frac{p^2}{\sqrt{n}} \max_{i,s,t} |[\boldsymbol{X}_i]_{st}| + \sqrt{\frac{2pq(q+1)}{n}}$$

Maximum softly-penalized likelihood

Consistency and asymptotic normality

To preserve ML asymptotics, scale $P(\boldsymbol{\theta})$ to make penalization asymptotically negligible

- Choose scaling factors c_1, c_2 to control $\|\nabla P(\boldsymbol{\theta})\|$ in terms of the rate of information accumulation r_n, which is such that

$$r_n^{-1}\left\{-\nabla\nabla^\top \ell(\boldsymbol{\theta})\right\} \xrightarrow{p} I(\boldsymbol{\theta}), \quad I(\boldsymbol{\theta}) = \mathcal{O}(1)$$

- For $c_1 = c_2 = 2\sqrt{p/n}$, $n = \sum_{i=1}^{k} n_i$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla P(\boldsymbol{\theta})\| \leq \frac{p^2}{\sqrt{n}} \max_{i,s,t} |[\boldsymbol{X}_i]_{st}| + \sqrt{\frac{2pq(q+1)}{n}}$$

✔ Under regularity conditions (Ogden, 2017), this bound is sufficient to establish consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}$ as long as $\max_{i,s,t} |[\boldsymbol{X}_i]_{st}| = O_p(n^{1/2})$

Simulation Study
Setup

Simulation based on the Culcita data from the motivating example

- Draw 10 000 independent samples of responses $\boldsymbol{Y}$ from the mixed effects logistic regression model of (3) at the ML estimate from the complete dataset
- Get ML, MSPL estimates using a 100-point adaptive Gauss-Hermite quadrature approximation to the log-likelihood
- Contrast with `bglmer` (Chung et al., 2013), a MPL estimator designed to prevent boundary estimates in (generalized) linear mixed models
  - Fixed effects penalty: independent t-priors and normal-priors
  - Variance components penalty: gamma-prior like penalty
- Compute summary statistics after discarding problematic estimates (Large parameter/standard error estimates, nonzero gradient)

Simulation Study

Results

MSPL

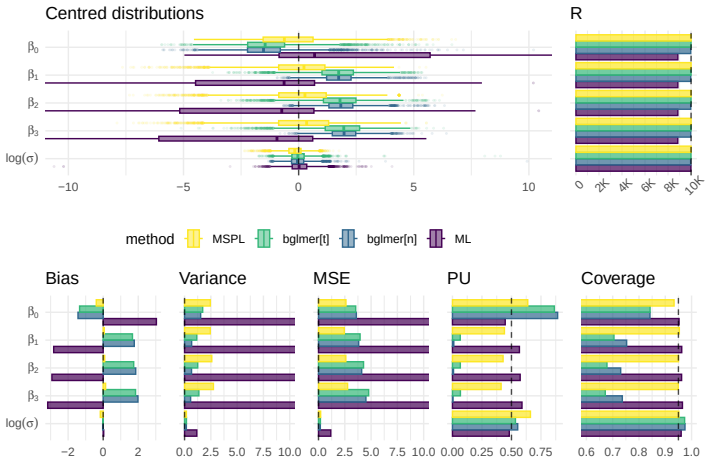
- small bias due to "soft penalization"
- accurate coverage

`bglmer`

- low variance due to excessive shrinkage
- large bias
- substantial undercoverage

ML

- performs poorly even when problematic estimates are thrown out



Thank you

WARWICK
THE UNIVERSITY OF WARWICK

Sterzinger P. and I. Kosmidis (2023). *Maximum softly-penalized likelihood for mixed effects logistic regression*. Statistics and Computing, 33(2).

References

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.

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.

Kosmidis, I. and D. Firth (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika 108*(1), 71–82.

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. (2014). On best practice optimization methods in R. *Journal of Statistical Software 60*(2), 1–14.

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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