Maximum softly-penalized likelihood

for mixed effects logistic regression

Ioannis Kosmidis

IKosmidis_

ikosmidis.com

ioannis.kosmidis@warwick.ac.uk

Professor of Statistics
University of Warwick

30 May 2023

Ca' Foscari University of Venice

Joint work with Philipp Sterzinger

Philipp

Sterzinger P, K. I (2023). Maximum softly-penalized likelihood for mixed effects logistic regression. Statistics and Computing, **33**, 53
DOI: 10.1007/s11222-023-10217-3

Outline

Outline

Mixed effects logistic regression

Data

Response vectors $y_1, \ldots, y_k$, $y_i = (y_{i1}, \ldots, y_{in_i})^\top \in \{0, 1\}^{n_i}$
Covariate matrices $V_1, \ldots, V_k$, with $V_i \in \Re^{n_i \times s}$

Model

$Y_{i1}, \ldots, Y_{in_i}$ conditionally independent with

$$
\begin{aligned}
Y_{ij} \mid u_i \sim \mathrm{Bern}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1 - \mu_{ij}} = \eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top u_i \\
U_i \sim \mathrm{N}(0_q, \Sigma) \quad (i = 1, \ldots, k; j = 1, \ldots, n_i)\,,
\end{aligned}
$$

$\beta \in \Re^p, \Sigma \in \Re^{q \times q}$

x_{ij}, z_{ij} are the jth row of matrices X_i, Z_i, respectively, constructed from columns of V_i

Maximum approximate likelihood estimation

Likelihood about β and Σ

$$L(\beta, \Sigma) \propto \det(\Sigma)^{-k/2} \prod_{i=1}^{k} \int_{\Re^q} \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}} e^{\frac{u_i^\top \Sigma^{-1} u_i}{2}} du_i$$

Maximum approximate likelihood estimator

$$\hat\beta, \hat\Sigma \in \arg\max_{\beta, \Sigma} L(\beta, \Sigma)$$

typically after numerically approximating $L(\beta, \Sigma)$

Certain data configurations and approximation methods result in estimates on the boundary of the parameter space:

- infinite or zero estimated variances, or, more generally, singular estimates of Σ
- estimates of β with infinite components

Motivating example

Culcita data (McKeon et al., 2012; worked examples of Bolker, 2015)

	Temporal 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

Coral-eating sea stars (Culcita) attacking coral harbouring protective symbionts (crabs, shrimp, both, none)

80 observations on whether predation was present (1) or not (0)

Complete randomized block design: 4 treatments, 10 temporal blocks, 2 repetitions per block-treatment combination

Motivating example

Culcita data (McKeon et al., 2012; worked examples of Bolker, 2015)

	Temporal 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

Predation suppressed when either crabs or shrimp present

Atypical observation: no predation with no symbionts in block 10

Motivating example

Model

Mixed effects logistic regression model with one random intercept per block to associate predation with treatment effects while accounting for heterogeneity between blocks

$$Y_{ij} \mid u_i \sim \mathrm{Bern}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = \beta_0 + \beta_{a(j)} + u_i$$

$$U_i \sim \mathrm{N}(0, \sigma^2) \quad (i = 1, \ldots, 10; j = 1, \ldots, 8)\,,$$

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

Maximum approximate likelihood

Set $\beta_1 = 0$ (“none” as reference category) for identifiability

Likelihood approximation: 100-point adaptive Gauss-Hermite quadrature

Remove atypical observation

Estimate $\beta, \log \sigma$ and asymptotic standard errors using the `optimx` (Nash, 2014) **R** package with methods “BFGS” and “CG”

	BFGS	CG
β_0	15.88	15.38
	(10.14)	(9.53)
β_2	-12.93	-12.47
	(9.15)	(8.56)
β_3	-14.81	-14.31
	(9.89)	(9.27)
β_4	-17.71	-17.16
	(10.70)	(10.06)
$\log \sigma$	2.31	2.28
	(0.64)	(0.62)

Large estimated standard errors are indicative of an almost flat approximate likelihood around the estimates

$\hat{\beta}$ actual value is $(+\infty, -\infty, -\infty, -\infty)$

Optimization procedures stopping early at different points in $\Re^5$ after having prematurely declared convergence

Boundary estimates in fixed-effects logistic regression: Detection

`detectseparation` R package (K. et al., 2022) provides linear programming methods of

- Schwendinger et al. (2021) (log-binomial regression)
- Konis (2007) (other binomial-response GLMs)

```
R> library("detectseparation")
R> culcita_fixed <- glm(predation ~ ttt, data = culcita_none, family = binomial,
+                       subset = !(block == 10 & predation == 0 & ttt == "none"))
R> coef(culcita_fixed)

(Intercept)    tttcrabs   tttshrimp     tttboth
   2.890372   -2.484907   -2.689701   -3.091042

R> update(culcita_fixed, method = detect_separation)

Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
(Intercept)    tttcrabs   tttshrimp     tttboth
          0           0           0           0
0: finite value, Inf: infinity, -Inf: -infinity
```

Boundary estimates in fixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Mean bias reducing adjusted scores

`brglm`, `brglm2`, `logistf` R packages

Firth (1993); K. and Firth (2021)

Median bias reducing adjusted score functions

`brglm2` R package

Kenne Pagui et al. (2017); K. et al. (2020)

Weakly informative priors

Gelman et al. (2008)

`brms` R package

. . .

Boundary estimates in mixed effects logistic regression: Detection

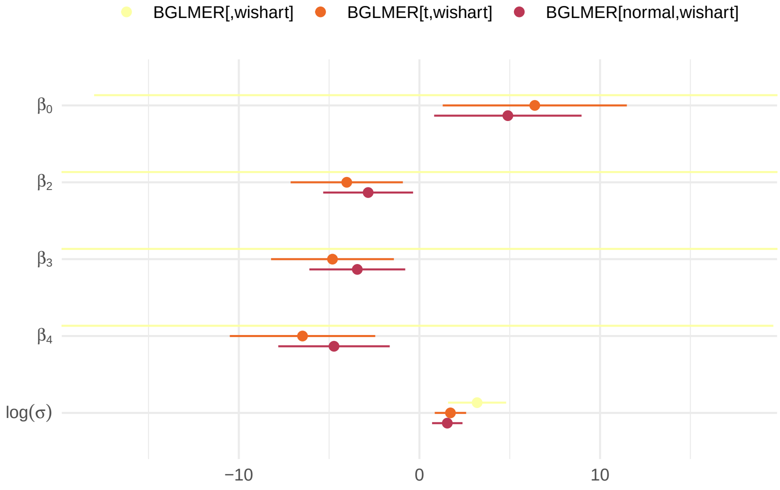
Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects to avoid boundary through prior-imposed shrinkage


BGLMER[,wishart]
BGLMER[t,wishart]
BGLMER[normal,wishart]
β_0
β_2
β_3
β_4
$\log(\sigma)$
−10
0
10

Contrasts

Parameterization with "none" as reference category (i.e. $\beta_1 = 0$)

$$\eta_{ij} = \beta_0 + \beta_{a(j)} + u_i$$

Parameterization with "both" as reference category (i.e. $\gamma_4 = 0$)

$$\eta_{ij} = \gamma_0 + \gamma_{a(j)} + u_i$$

So,

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

Natural to expect the estimation method to deliver estimators $\tilde{\gamma}$ and $\tilde{\beta}$ that respect the above identities, i.e. $\tilde{\gamma} = C\tilde{\beta}$

Contrasts

ML

	β	$C\beta$	γ
β_0	15.88		
	(10.14)		
β_2	−12.93		
	(9.15)		
β_3	−14.81		
	(9.89)		
β_4	−17.71		
	(10.70)		
$\log \sigma$	2.31	2.31	2.31
	(0.64)	(0.64)	(0.64)
γ_0		−1.82	−1.82
		(3.92)	(3.92)
γ_1		17.71	17.74
		(10.70)	(10.75)
γ_2		4.78	4.78
		(3.07)	(3.08)
γ_3		2.89	2.89
		(2.27)	(2.27)

BGLMER[normal,wishart]

	β	$C\beta$	γ
β_0	4.90		
	(2.08)		
β_2	−2.84		
	(1.27)		
β_3	−3.44		
	(1.35)		
β_4	−4.73		
	(1.57)		
$\log \sigma$	1.54	1.54	1.66
	(0.43)	(0.43)	(0.44)
γ_0		0.17	0.57
		(1.83)	(2.07)
γ_1		4.73	5.75
		(1.57)	(1.88)
γ_2		1.89	1.26
		(1.32)	(1.32)
γ_3		1.29	0.56
		(1.27)	(1.28)

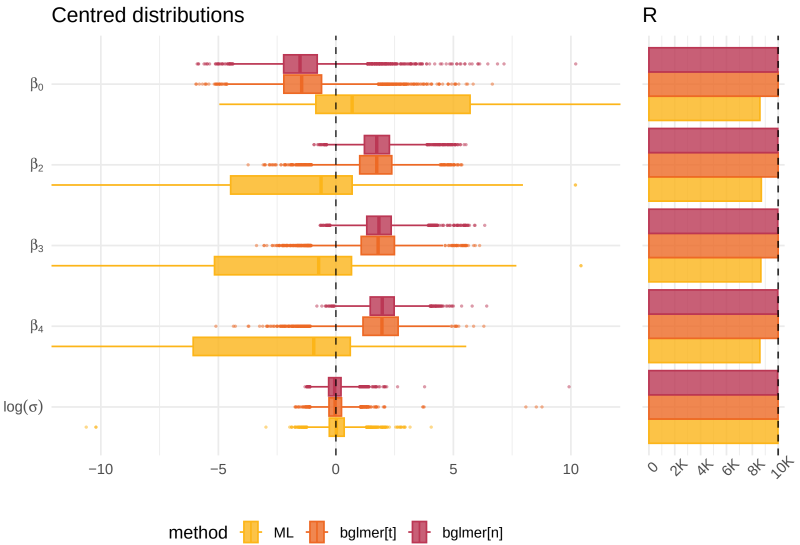
Simulation study

parameter	value
β_0	5.01
β_2	-3.75
β_3	-4.36
β_4	-5.55
$\log \sigma$	1.26

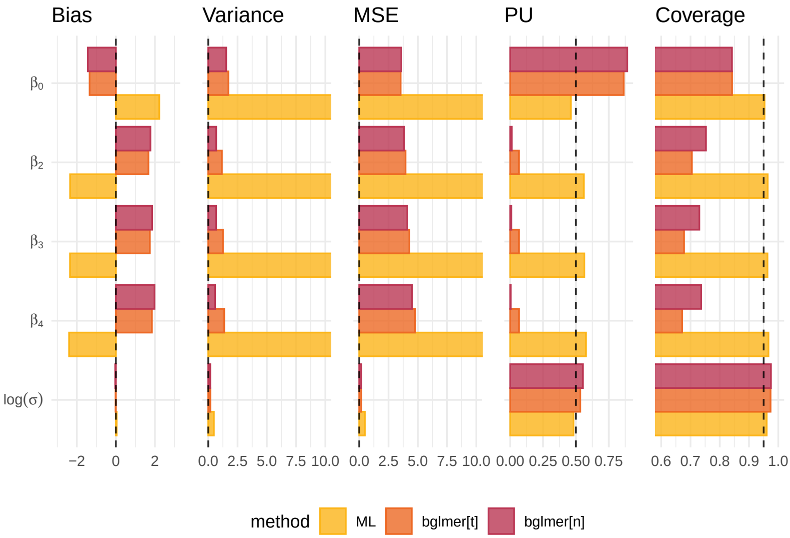
$R = 10000$ independent samples

Ignore samples where any of the

- $|$estimated gradient components$|$ is $> 10^{-3}$, or
- $|$estimates$|$ or estimated standard errors is > 30



Centred distributions
R
β0
β2
β3
β4
log(σ)
−10
−5
0
5
10
0
2K
4K
6K
8K
10K
method
ML
bglmer[t]
bglmer[n]


Bias
Variance
MSE
PU
Coverage
β0
β2
β3
β4
log(σ)
−2
0
2
0.0
2.5
5.0
7.5
10.0
0.00
0.25
0.50
0.75
0.6
0.7
0.8
0.9
1.0
method
ML
bglmer[t]
bglmer[n]

Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects

Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects

What prior?

Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects

What prior?

Default prior for random effects variance is $3 \log \sigma / 2$
$\longrightarrow$
no guarding against infinite variance estimates

Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects

What prior?

Default prior for random effects variance is $3 \log \sigma / 2$
$\longrightarrow$
no guarding against infinite variance estimates

How much shrinkage for optimal frequentist properties?

Boundary estimates in mixed-effects logistic regression: Alternative estimators

Estimates that are expected not to be on the boundary

Weakly informative priors for covariance estimation

Chung et al. (2013, 2015)

`bglmer` R package, which also adds support for default prior penalties to the fixed effects

What prior?

Default prior for random effects variance is $3 \log \sigma / 2$
$\longrightarrow$
no guarding against infinite variance estimates

How much shrinkage for optimal frequentist properties?

Invariance to simple contrasts or scaling of covariates?

Outline

Maximum softly-penalized likelihood

Setup:

Penalize (approximate) log-likelihood with penalty that diverges to $-\infty$ when we approach the boundary of the parameter space

Ensure fixed effects estimates are equivariant under linear transformations of the model parameters

Make penalization "soft"-enough for the MSPL estimator to have the same optimal asymptotic properties expected by the ML estimator

Definitions and notation

Model parameters

Let $\theta = (\beta^\top, \psi^\top)^\top$, where

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

with l_{ij} $(i > j)$ the (i,j)th element of L, from $\Sigma = s(\psi) = LL^\top$

MPL estimator

For $\ell(\theta) = \log L(\beta, s(\psi))$ and penalty function $P(\theta)$, define

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

Composite penalties

Penalties of the form

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

$c_1 > 0$, $c_2 > 0$

$P_{(f)}(\beta)$ is the unscaled fixed effects penalty

$P_{(v)}(\psi)$ is the unscaled variance components penalty

$P_{(f)}(\beta)$ and $P_{(v)}(\psi)$

Fixed effects penalty

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

$$P_{(f)}(\beta) = \frac{1}{2} \log \det \sum_{i=1}^{k} X_i^\top W_i X_i$$

X_i collects the fixed effects covariates of cluster i

W_i is diagonal with jth diagonal element $\exp(x_{ij}^\top \beta)/\{1 + \exp(x_{ij}^\top \beta)\}^2$

Variance components penalty

Composition of negative Huber loss functions on the components of ψ

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

Non-boundary estimates

Let $\theta(r)$, $r \in \Re$, be a path in Θ such that $\lim_{r\to\infty} \theta(r) \in \partial\Theta$

$P_{(f)}(\beta) \to -\infty$ if at least one β diverges (K. and Firth, 2021, Th. 1)
$P_{(v)}(\psi) \to -\infty$ if at least one ψ diverges (by construction)

So, for any $c_1 > 0$ and $c_2 > 0$, $P(\theta(r)) \to -\infty$

Given that also $\ell(\theta)$ is bounded from above and is not $-\infty$ for all $\theta \in \Theta$, it can be shown that

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

Equivariance under linear transformations of fixed effects

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

If the fixed-effects penalty behaves like the log-likelihood under linear transformations $\gamma = C\beta$, then the MPL estimators are equivariant

For any known $C \in \Re^{p \times p}$, $P_{(f)}(C\beta) = P_{(f)}(\beta) - \log\det C$. So, $\tilde{\gamma} = C\tilde{\beta}$

Soft penalties for consistency and asymptotic normality I

Information accumulation and choice of c_1 and c_2

Choose scaling factors $c_1, c_2 > 0$ to control $\|\nabla P(\theta)\|$ in terms of the rates of information accumulation about the model parameters

Variances of fixed-effects linear predictors

Using the delta method, the square root of the average of the approximate variances of $x_{ij}^\top \hat\beta$ at $\beta = 0$ is $2\sqrt{p/n}$, $n = \sum_{i=1}^k n_i$

For $c_1 = c_2 = 2\sqrt{p/n}$

$$\|\nabla P(\theta)\| \le \frac{p^2}{\sqrt{n}} \max_{i,s,t} |[X_i]_{st}| + \sqrt{\frac{2pq(q+1)}{n}} \qquad (1)$$

Soft penalties for consistency and asymptotic normality II

Consistency and asymptotic normality

If $\max_{i,s,t} |[X_i]_{st}| = O_p(n^{1/2})$ in (1), then similar arguments to those in Ogden (2017) guarantee $\tilde\theta$'s consistency and asymptotic normality

$\max_{i,s,t} |[X_i]_{st}| = O_p(n^{1/2})$ is reasonable in practice. E.g.

- covariates encoding factors and interactions of those
- sub-Gaussian random variables with variance proxy σ^2

Outline

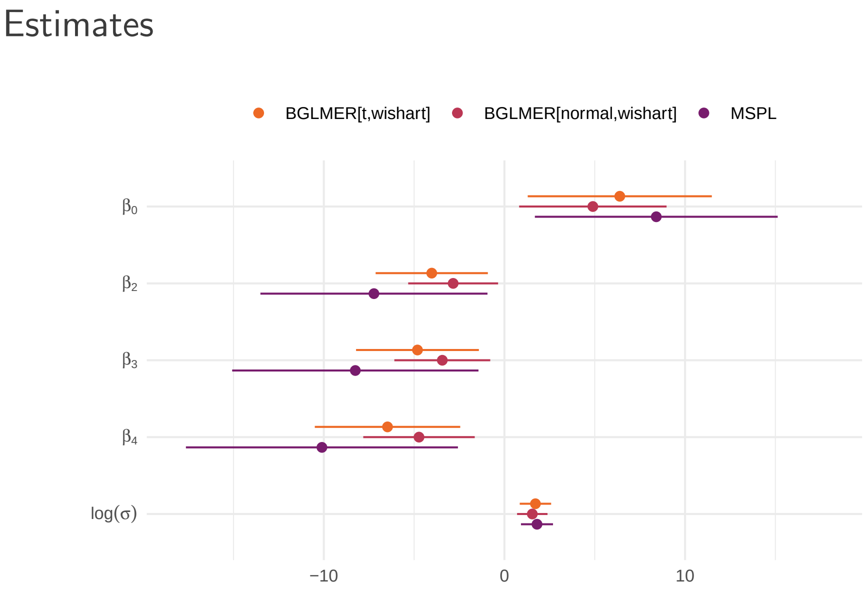
Contrasts

ML

	β	$C\beta$	γ
β_0	15.88		
	(10.14)		
β_2	−12.93		
	(9.15)		
β_3	−14.81		
	(9.89)		
β_4	−17.71		
	(10.70)		
$\log\sigma$	2.31	2.31	2.31
	(0.64)	(0.64)	(0.64)
γ_0		−1.82	−1.82
		(3.92)	(3.92)
γ_1		17.71	17.74
		(10.70)	(10.75)
γ_2		4.78	4.78
		(3.07)	(3.08)
γ_3		2.89	2.89
		(2.27)	(2.27)

MSPL

	β	$C\beta$	γ
β_0	8.41		
	(3.43)		
β_2	−7.23		
	(3.21)		
β_3	−8.26		
	(3.48)		
β_4	−10.10		
	(3.84)		
$\log\sigma$	1.80	1.80	1.80
	(0.45)	(0.45)	(0.45)
γ_0		−1.70	−1.70
		(2.46)	(2.46)
γ_1		10.10	10.10
		(3.84)	(3.84)
γ_2		2.88	2.88
		(1.86)	(1.86)
γ_3		1.85	1.85
		(1.60)	(1.60)



Estimates
BGLMER[t,wishart]
BGLMER[normal,wishart]
MSPL
β0
β2
β3
β4
log(σ)
−10
0
10

Simulation study

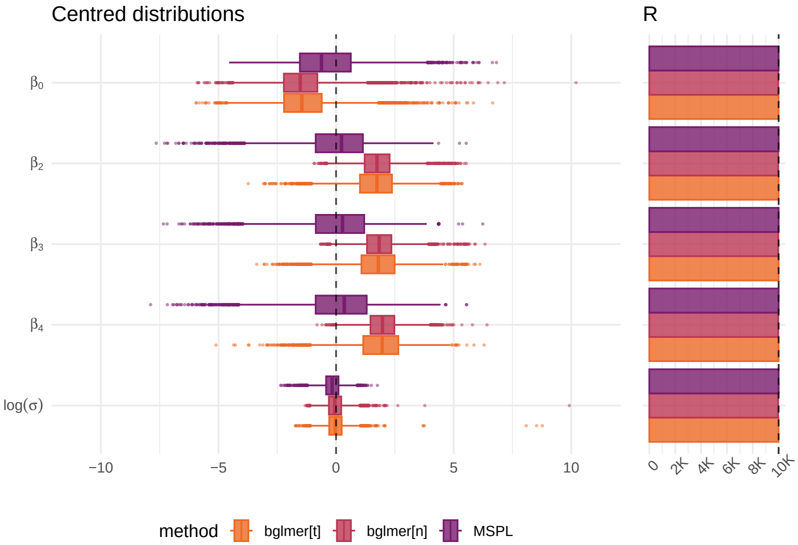
parameter	value
β_0	5.01
β_2	-3.75
β_3	-4.36
β_4	-5.55
$\log \sigma$	1.26

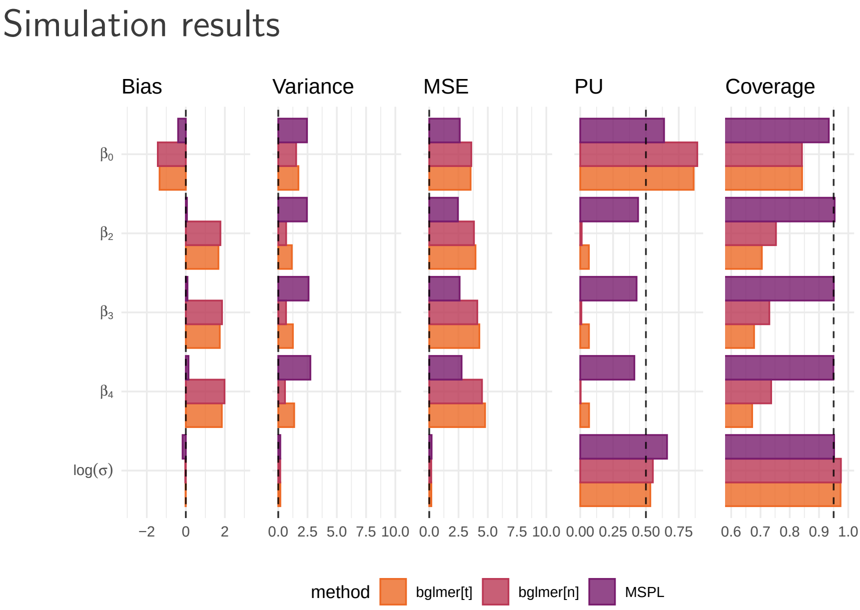
$R = 10000$ independent samples

Ignore samples where any of the

- $|$estimated gradients$|$ is $> 10^{-3}$, or
- $|$estimates$|$ or estimated standard errors is > 30

Simulation results





Simulation results
Bias
Variance
MSE
PU
Coverage
β0
β2
β3
β4
log(σ)
−2
0
2
0.0
2.5
5.0
7.5
10.0
0.0
2.5
5.0
7.5
10.0
0.00
0.25
0.50
0.75
0.6
0.7
0.8
0.9
1.0
method
bglmer[t]
bglmer[n]
MSPL

Extreme fixed effects

Model

$$Y_{ij} \mid u_i \sim \text{Bern}(\mu_{ij}) \quad \text{with} \quad \log \frac{\mu_{ij}}{1-\mu_{ij}} = \eta_{ij} = x_{ij}^\top \beta + u_i$$

$$U_i \sim \text{N}(0, 9) \quad (i = 1, \ldots, 5; j = 1, \ldots, n)\,,$$

Covariates

$x_{i1} = 1$, $x_{i2} \sim \text{N}(0, 1)$, $x_{i3} \sim \text{Bern}(1/2)$, $x_{i4} \sim \text{Bern}(1/2)$, $x_{i5} \sim \textit{Exp}(1)$

Simulation setup

$\beta = (1, -0.5, \lambda, 0.25, -1)$

For each $n \in \{50, 100, 200\}$, simulate covariates

For each $n \in \{50, 100, 200\}$ and $\lambda \in (-10, 10)$, draw 100 independent response vectors from the model

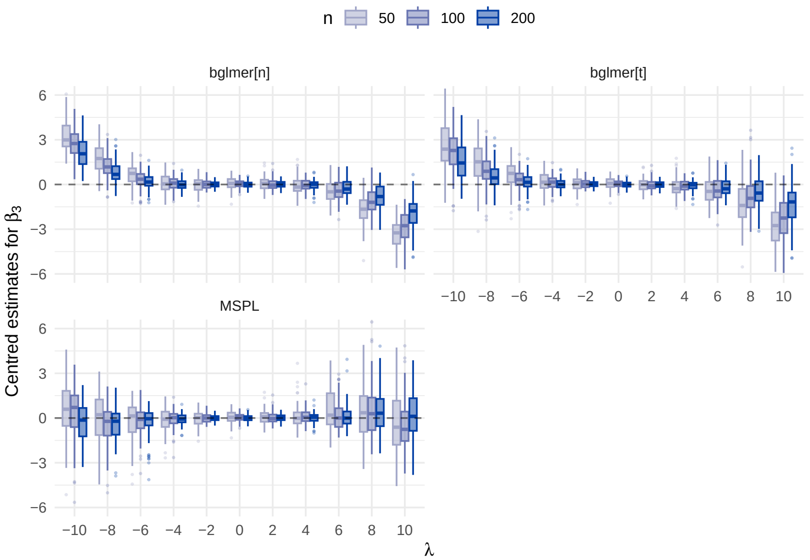
Extreme fixed effects

Samples where

- |estimated partial derivative| for β_3 is $> 10^{-3}$ or
- |estimate| or estimated standard error for β_3 is > 30

		λ										
		-10	-8	-6	-4	-2	0	2	4	6	8	10
ML	$n = 50$	73	46	17	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[n]	$n = 50$	1	0	0	0	0	1	0	0	0	0	0
	$n = 100$	0	0	1	0	1	0	0	0	1	0	0
	$n = 200$	0	0	0	0	0	0	0	0	0	0	0
bglmer[t]	$n = 50$	1	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	0
MSPL	$n = 50$	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

Extreme fixed effects



Outline

Remarks

Soft penalization restores and preserves the optimal asymptotic properties expected by ML, while ensuring no boundary estimates.

Concept is far more general and can be adopted in other GLMMs with degenerate estimates (other links, nominal/ordinal responses).

Composite negative Huber loss penalty can be adapted to prevent singular variance-covariance estimates more generally.

Reduced-bias M-estimation methodology (K. and Lunardon, 2021) readily applies to MSPL estimators.

Sterzinger P, K. I (2023). Maximum softly-penalized likelihood for mixed effects logistic regression. Statistics and Computing, **33**, 53
DOI: 10.1007/s11222-023-10217-3

A soft composite penalty

$$P(\theta) = 2\sqrt{p/n}\left\{P_{(f)}(\beta) + P_{(v)}(\psi)\right\}$$

$$P_{(f)}(\beta) = \frac{1}{2}\log\det\sum_{i=1}^{k} X_i^\top W_i X_i$$

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

IKosmidis_ ioannis.kosmidis@warwick.ac.uk www.ikosmidis.com

References I

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., 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.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

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.

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

K., I., E. C. Kenne Pagui, and N. Sartori (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing (to appear) 30*, 43–59.

K., I. and N. Lunardon (2021). Empirical bias-reducing adjustments to estimating functions. *arXiv:2001.03786*.

K., I., D. Schumacher, and F. Schwendinger (2022). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.3.

Kenne Pagui, E. C., A. Salvan, and N. Sartori (2017). Median bias reduction of maximum likelihood estimates. *Biometrika 104*(4), 923–938.

References II

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

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.

Schwendinger, F., B. Grün, and K. Hornik (2021). A comparison of optimization solvers for log binomial regression including conic programming. *Computational Statistics*.

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