mixed models: challenges and frontiers

Ben Bolker

18 November 2025

Table of contents

- Inference details
- Troubleshooting
- Structured covariances
- Big data
- Challenges and open questions

Inference details

Confidence intervals/p-values

- as with GLMM fitting, methods range from “easy, less accurate” to “computationally expensive, more accurate”
  - Wald approximations
  - Likelihood ratio test/Profile likelihood
  - Bootstrap
  - MCMC (Bayesian) ...
- easy methods work better as number of observations *and* number of clusters grows
- easy methods work better for fixed effects than for RE variances

Wald approximation: hypothesis tests

- assumes log-quadratic likelihood surface
  (exact for linear models)
- output of `summary()` (and `car::Anova()`, `afex::anova()`)
- contrasts/post-hoc tests: `emmeans`
- especially unreliable for random-effects (co)variances/correlations
- confidence intervals: `confint(., method = "Wald")` (fixed only)

Degrees of freedom

- account for uncertainty in variance estimation (Z vs t)
- matters when number of groups is ‘small’ (<50?)
- level-counting: `nlme::lme()`, `R/calcDenDF.R`
- `lmerTest/afex`; Satterthwaite:

```
##             Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)   251.04       8.72 12.12   28.80  1.6e-12
## Days            9.69       2.09 12.26    4.63  0.00055
```

- Kenward-Roger: only applicable to REML fits (LMMs?); also adjusts variances

```
##             Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)   251.04       8.72 11.99   28.79    2e-12
## Days            9.69       2.10 11.86    4.62  0.00061
```

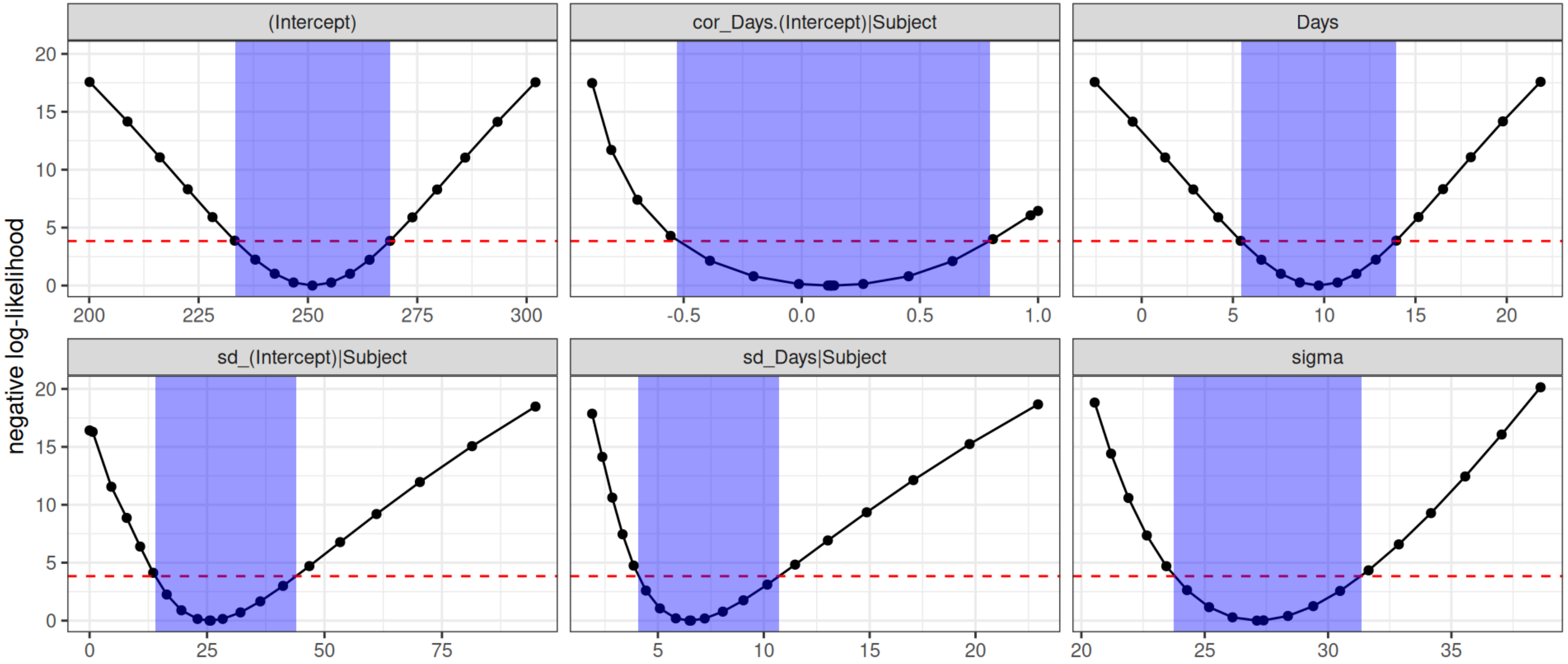
Likelihood ratio test/profile confidence intervals

- individual parameters: `profile()`, `confint()`
- fit pairwise models and use `anova()` to compare
- `drop1`, `afex::anova()`

```
##                                 2.5 %  97.5 %
## sd_(Intercept)|Subject         13.934  43.987
## cor_Days.(Intercept)|Subject   -0.528   0.796
## sd_Days|Subject                 4.072  10.732
## sigma                          23.757  31.341
## (Intercept)                   233.352 268.740
## Days                            5.433  13.929
```

Likelihood profiles

Why did `confint()` take so long?



Likelihood ratio tests of variances are problematic

- theory of LRT fails when H_0 is “on the boundary”
- e.g. testing whether variance is 0
- simplest cases, conservative (p-vals are 2X nominal)
- do you really need to hypothesis-test random effects?
- bootstrapping …

Bootstrap

- Thai et al. (2013)
- *nonparametric*: resampling with replacement (`lmeresampler` package)
- straightforward for nested models
- hard for crossed models: *nonparametric* bootstrapping
    - simulate from fitted model, refit, extract estimates
    - `confint(., method = "boot")`, `bootMer()`, `pbkrtest` package
- slow! (some methods are parallelized)
- estimate CIs on any quantity, e.g. ratios of variances

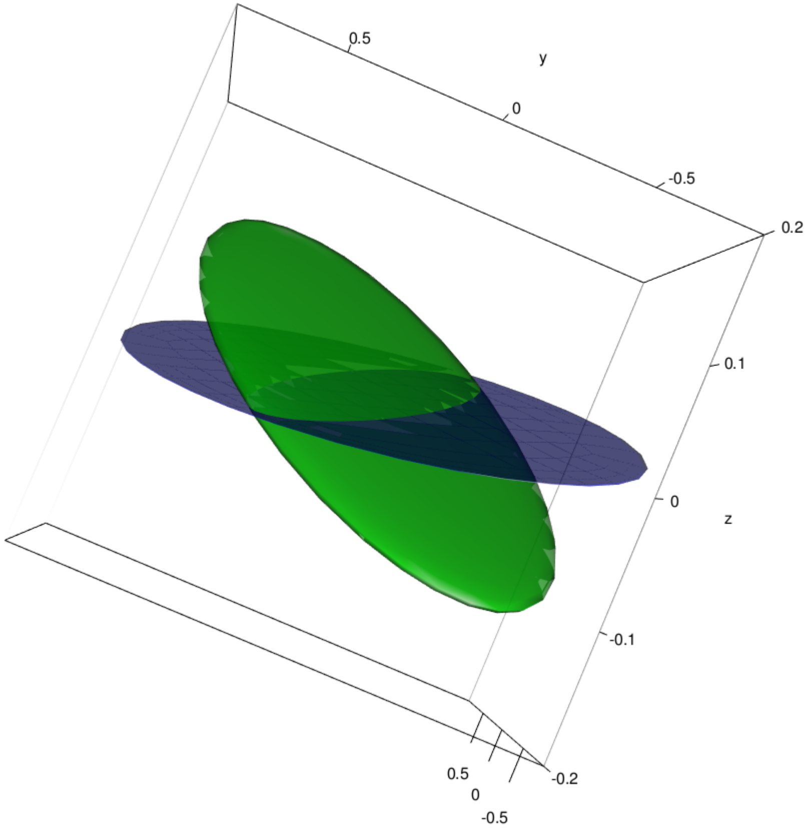
Predictive simulation

- `simulate()` new data from fitted model
- parameters fixed at best estimates (unlike nonpar bootstrap)
- can either use ‘estimated’ conditional modes or draw new random values
- does *not* include parameter uncertainty

Troubleshooting

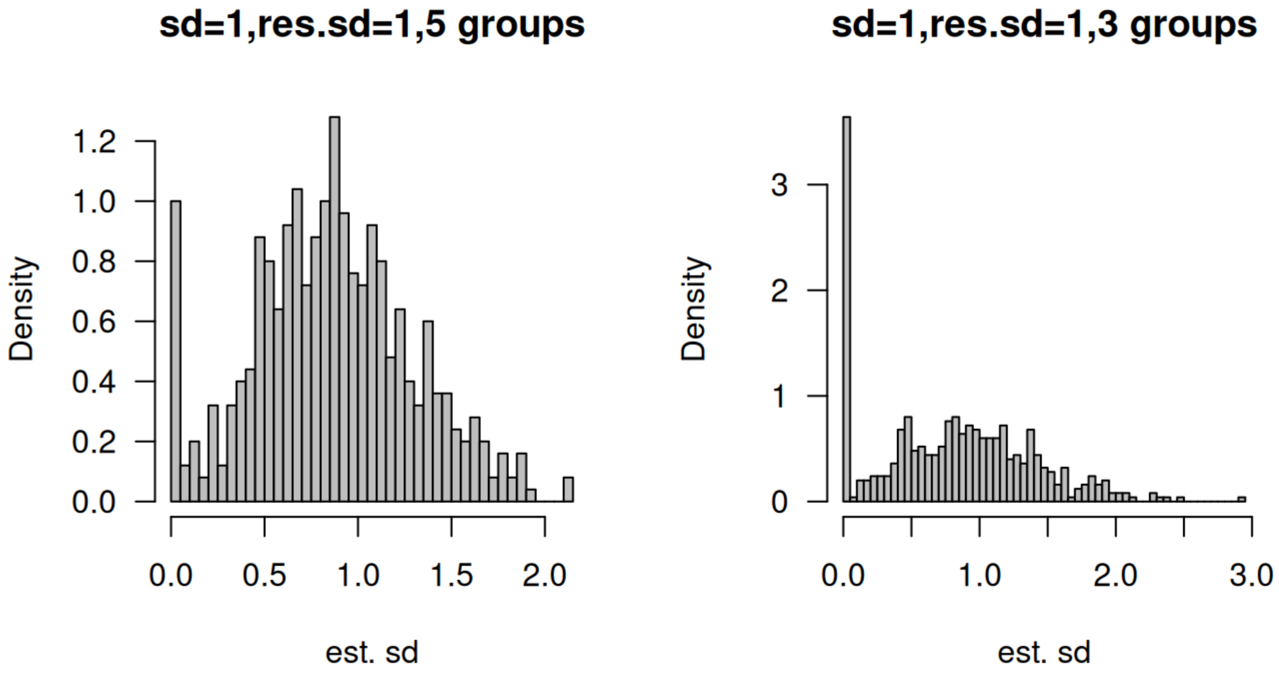
What is a “singular fit”?

- non-positive-definite covariance matrix
- zero variance (simple case)
- ±1 correlation (next simplest case)
- hard to recognize in the general case!
- some linear combination(s) of random-effect variables have zero variance



Why are fits singular?

Essentially, because the *observed* among-group variation is less than the *expected* among-group variation $\left(=\frac{\sigma^2_{\text{among}}+\sigma^2_{\text{within}}}{n}\right)$. More generally, because *some* dimension of the variance-covariance matrix has zero extent …



What to do about singular fits?

- ignore them, or drop terms
- switch from random to fixed effects
- regularize/add priors
- simplify the model
- full Bayesian treatment

Ignoring/dropping terms

- Zero variances/singular fits *are not a mistake*; they're the best fit to the data
- Dropping a (full) singular term gives the same results/predictions/etc.
    - It does narrow confidence intervals (if using profile CIs)

Switch from random to fixed effects

- especially for top-level clusters with few levels
- admit you can't estimate the variance (RE variance → ∞)
  - more conservative than dropping/pooling (RE variance = 0)
- e.g. *Arabidopsis*: region (3) / population (9) / genotype (24), use ~ `reg + (1|(reg:population)/genotype)`

Regularize/add priors

- Bayesians don’t care about singular fits
  (they look at the whole posterior distribution, not just the MLE/mode)
- quasi-Bayesian approach: *regularize*
- add a penalty (== prior) to keep estimates away from zero/boundaries
- Chung et al. (2013): add a weak prior that prevents singularity
- `blme`, `glmmTMB` packages
- useful in other cases (complete separation in logistic models)

Simplifying random effect terms

- drop some varying effects
  - e.g. random intercepts and slopes → random intercepts only
- set correlations among parameters to zero (diagonal covariance matrix)
- set all correlations among parameters equal (*compound symmetric* model)
- *reduced rank* models

Simplifying covariance matrices

- Scandola et al. (2024)
- Suppose factor `f` (5 levels) varies within groups `g`
- `(1+f|g)`: $(n(n+1))/2 = (5 \times 6)/2 = 15$ parameters)

$$\begin{bmatrix} \sigma^2_{\{g|1\}} & . & . & . & . \\ \sigma_{\{g|1\},\{g|f_{21}\}} & \sigma^2_{\{g|f_{21}\}} & . & . & . \\ \sigma_{\{g|1\},\{g|f_{31}\}} & \sigma_{\{g|f_{21}\},\{g|f_{31}\}} & \sigma^2_{\{g|f_{31}\}} & . & . \\ \sigma_{\{g|1\},\{g|f_{41}\}} & \sigma_{\{g|f_{21}\},\{g|f_{41}\}} & \sigma_{\{g|f_{31}\},\{g|f_{41}\}} & \sigma^2_{\{g|f_{41}\}} & . \\ \sigma_{\{g|1\},\{g|f_{51}\}} & \sigma_{\{g|f_{21}\},\{g|f_{51}\}} & \sigma_{\{g|f_{31}\},\{g|f_{51}\}} & \sigma_{\{g|f_{41}\},\{g|f_{51}\}} & \sigma^2_{\{g|f_{51}\}} \end{bmatrix}$$

Sum-to-zero contrasts

- change from *treatment* to *sum-to-zero* contrasts
- coefficients are means of each level, not baseline/differences from baseline

$$\begin{bmatrix} \sigma^2_{\{g|f_1\}} & \cdot & \cdot & \cdot & \cdot \\ \sigma_{\{g|f_1\},\{g|f_2\}} & \sigma^2_{\{g|f_2\}} & \cdot & \cdot & \cdot \\ \sigma_{\{g|f_1\},\{g|f_3\}} & \sigma_{\{g|f_2\},\{g|f_3\}} & \sigma^2_{\{g|f_3\}} & \cdot & \cdot \\ \sigma_{\{g|f_1\},\{g|f_4\}} & \sigma_{\{g|f_2\},\{g|f_4\}} & \sigma_{\{g|f_3\},\{g|f_4\}} & \sigma^2_{\{g|f_4\}} & \cdot \\ \sigma_{\{g|f_1\},\{g|f_5\}} & \sigma_{\{g|f_2\},\{g|f_5\}} & \sigma_{\{g|f_3\},\{g|f_5\}} & \sigma_{\{g|f_4\},\{g|f_5\}} & \sigma^2_{\{g|f_5\}} \end{bmatrix}$$

Complex random effects

- replace `(1+f|g)` with `(1|g/f)`

$$\Sigma = \begin{bmatrix} \sigma^2 & . & . & . & . \\ \rho\sigma^2 & \sigma^2 & . & . & . \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & . & . \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & . \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

- where $\sigma^2 = \sigma_f^2 + \sigma_g^2$, $\rho = \frac{\sigma_g^2}{\sigma^2}$
- $\rho > 0$; all variances identical
- `cs()` (glmmTMB), `corCompSymm` (nlme)

Diagonal covariance matrices

- `(1+x+y+z||b)` or `diag(1+x+y+z|b)`
    - or `(1|b) + (0+x|b) + ...`
- `lme4` version **only works properly for continuous predictors**
- `afex::mixed` can do this
- contrasts/centering matters!
- n instead of $n(n+1)/2$ parameters

Reduced-rank covariance matrices

- ‘factor analytic’ or ‘reduced rank’
- d-dimensional covariance matrix in a p-dimensional space
- $\Sigma = \Lambda D \Lambda^\top$ (Λ is the *factor loading matrix*)
- $d \cdot p - \frac{d(d-1)}{2}$ parameters (linear in p)
- available in R: `glmmTMB`, `sommer`, `lme4breeding`: ASReml

Convergence failures ☠

- convergence failures are common
- what do they really mean? how to fix them? when can they be ignored?
- **approximate** test that gradient=0 and curvature is correct
- scale and center predictors; simplify model
- use `?allFit` to see whether different optimizers give sufficiently similar answers
  - `$fixef`, etc.: are answers sufficiently similar?
  - `$llik`: how similar is goodness-of-fit?

Which model to use?

- Barr et al. (2013): “keep it maximal”; reduce until no more convergence failures
- Bates et al. (2015), Matuschek et al. (2017): more parsimony
- R. Kliegl: best non-singular
    - make singular terms diagonal
    - drop elements with small variances
    - try full (unstructured) model with only these elements
- brute-force: non-singular, best AIC (Moritz et al. 2023)
- inspect principal components of covariance matrix? (`rePCA`)

Structured covariances

Spatial/temporal

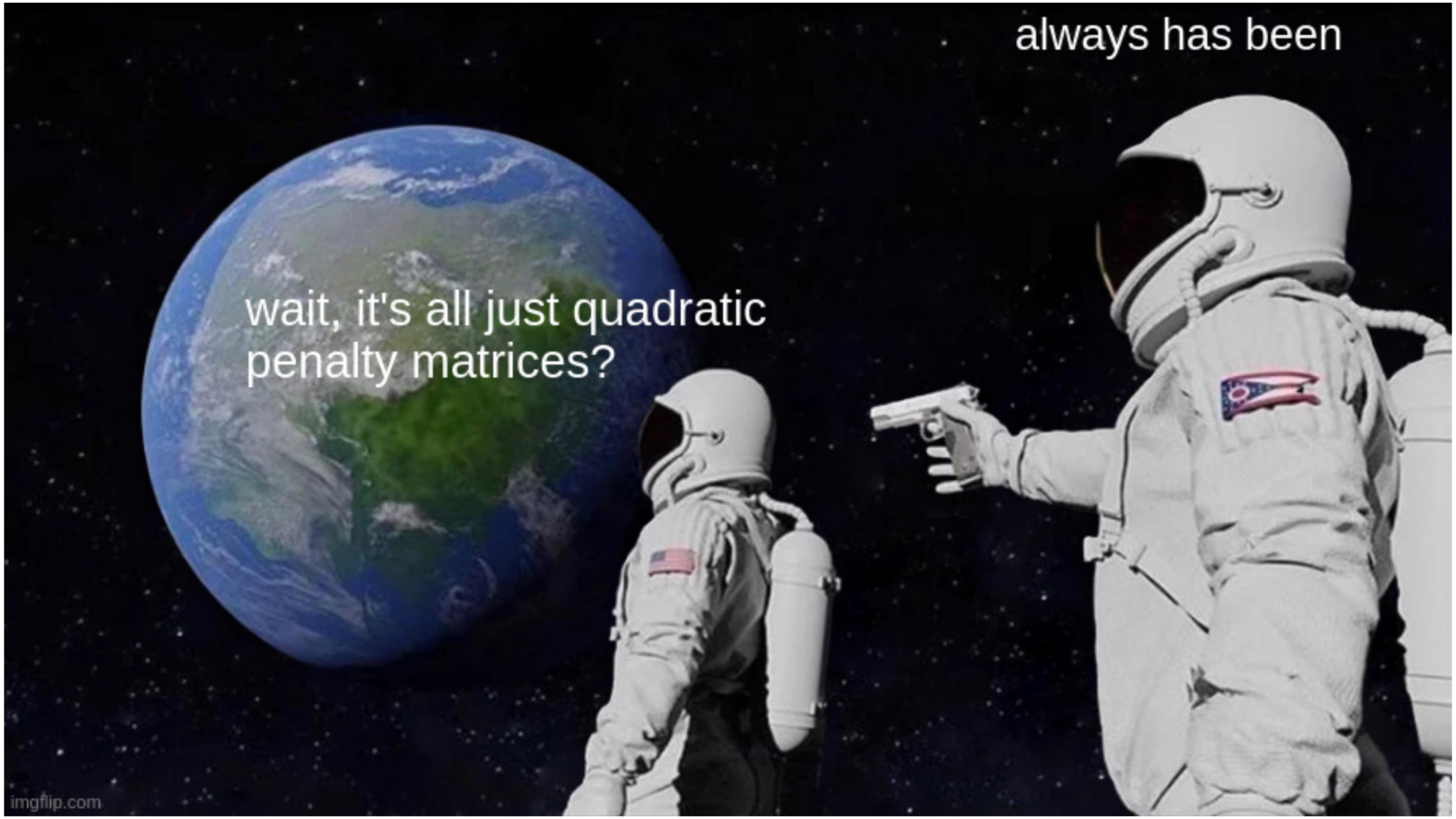
- AR1 (autoregressive order-1): correlations ρ, ρ^2, ρ^3 … for lagged pairs in space/ time
- irregular time/1D space: Ornstein-Uhlenbeck/CAR1 ($\rho = \phi^{-\Delta t}$)
- geostatistical (Gaussian-process) correlation models: exponential, Matérn …
- `glmmTMB`, `mgcv`, `INLA`, `spaMM`

Smooth terms

- can use mixed-model/latent-variable frameworks to fit smooth *functional* terms
- additive models (penalized regression): add penalty $\sigma^2 b S b^\top$
- exactly equivalent to the random-effects component in a mixed model
- Wahba (1990); Wood (2017); Hefley et al. (2017); Pedersen et al. (2018)
- `mgcv`, `gamm4`, `glmmTMB`

‘Old’ and ‘new’ random effects (Hodges 2016)

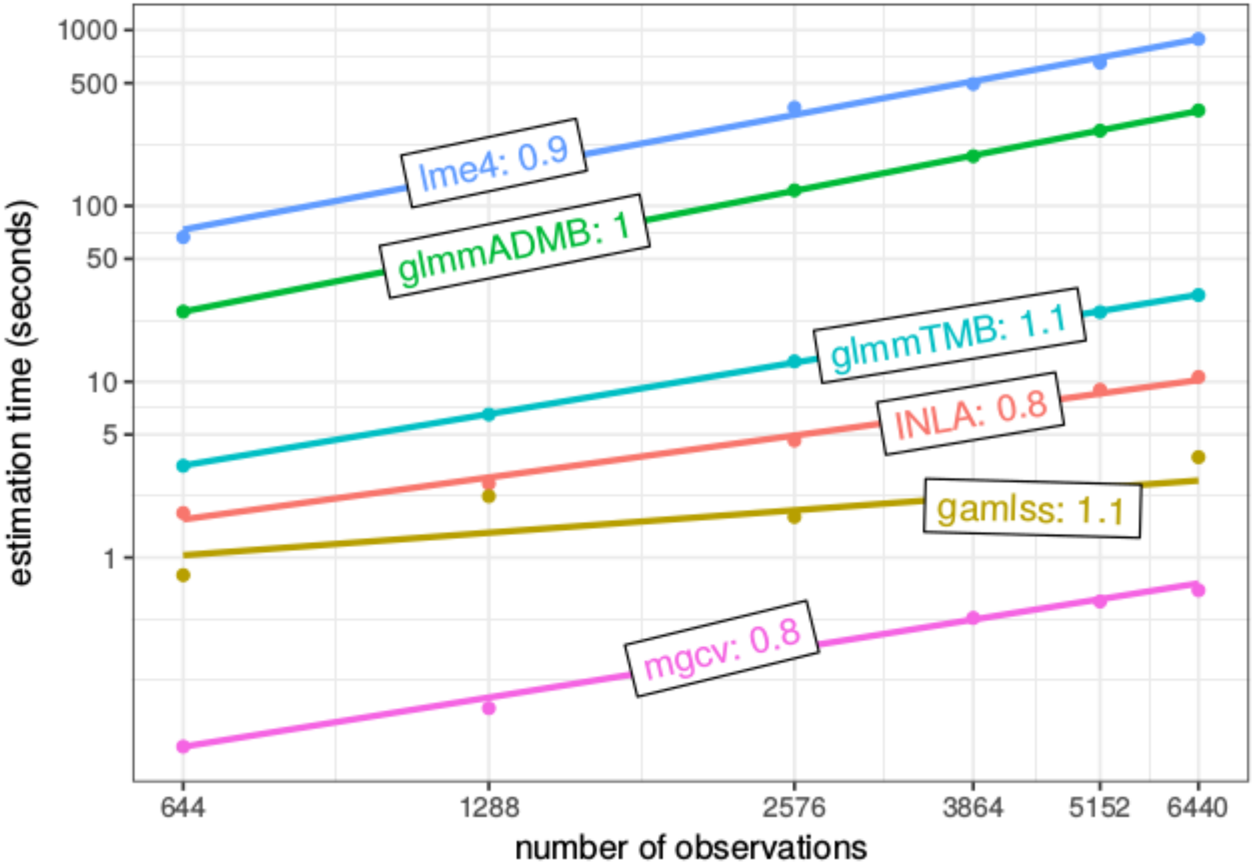
- do random effects/latent variables represent differences among exchangeable groups …
- … or spatial/temporal/functional structure? (Wood 2017)
- prediction/inference: population level or the cluster level (Vaida et al. 2005) ?
- counting parameters?


always has been
wait, it's all just quadratic penalty matrices?
imgflip.com

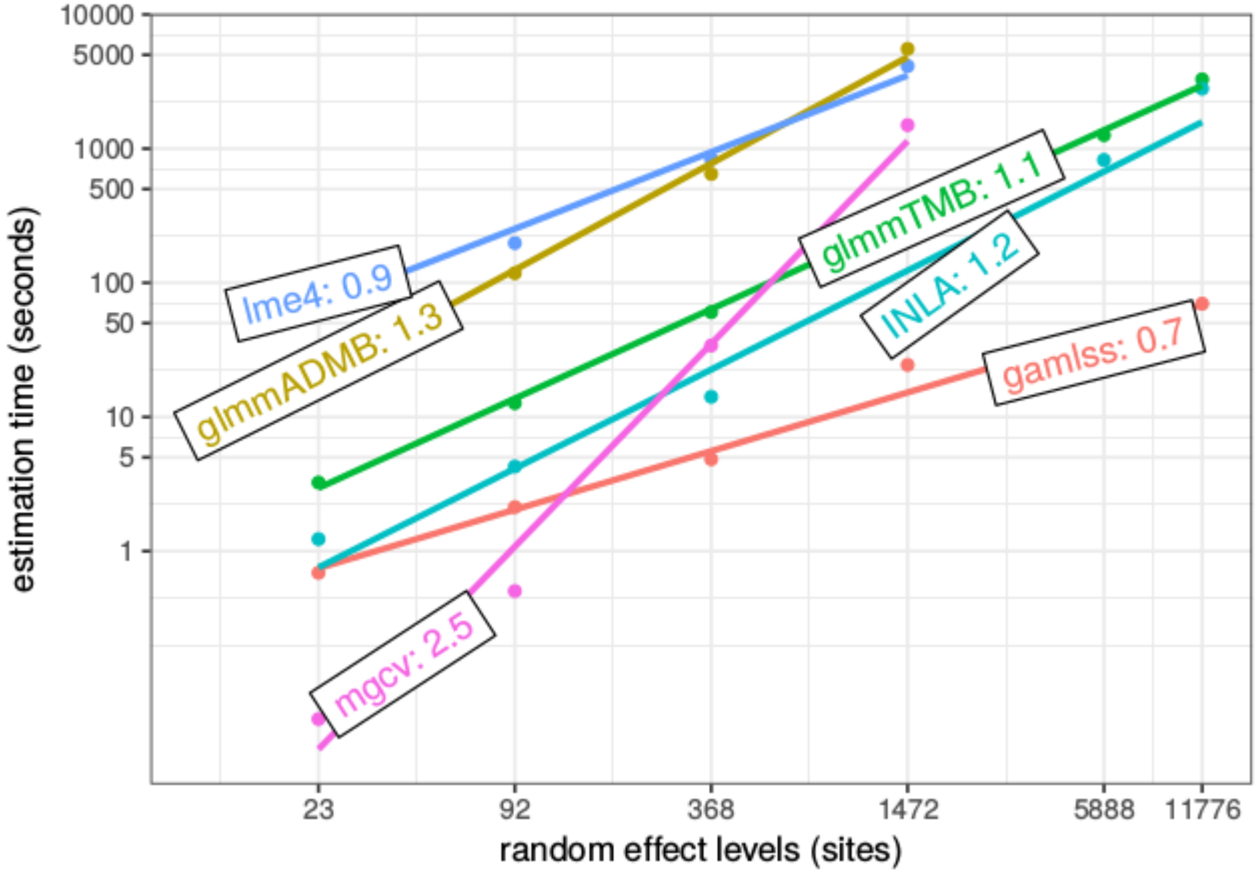
Big data?

The need for speed

Brooks et al. (2017) (negative binomial model)

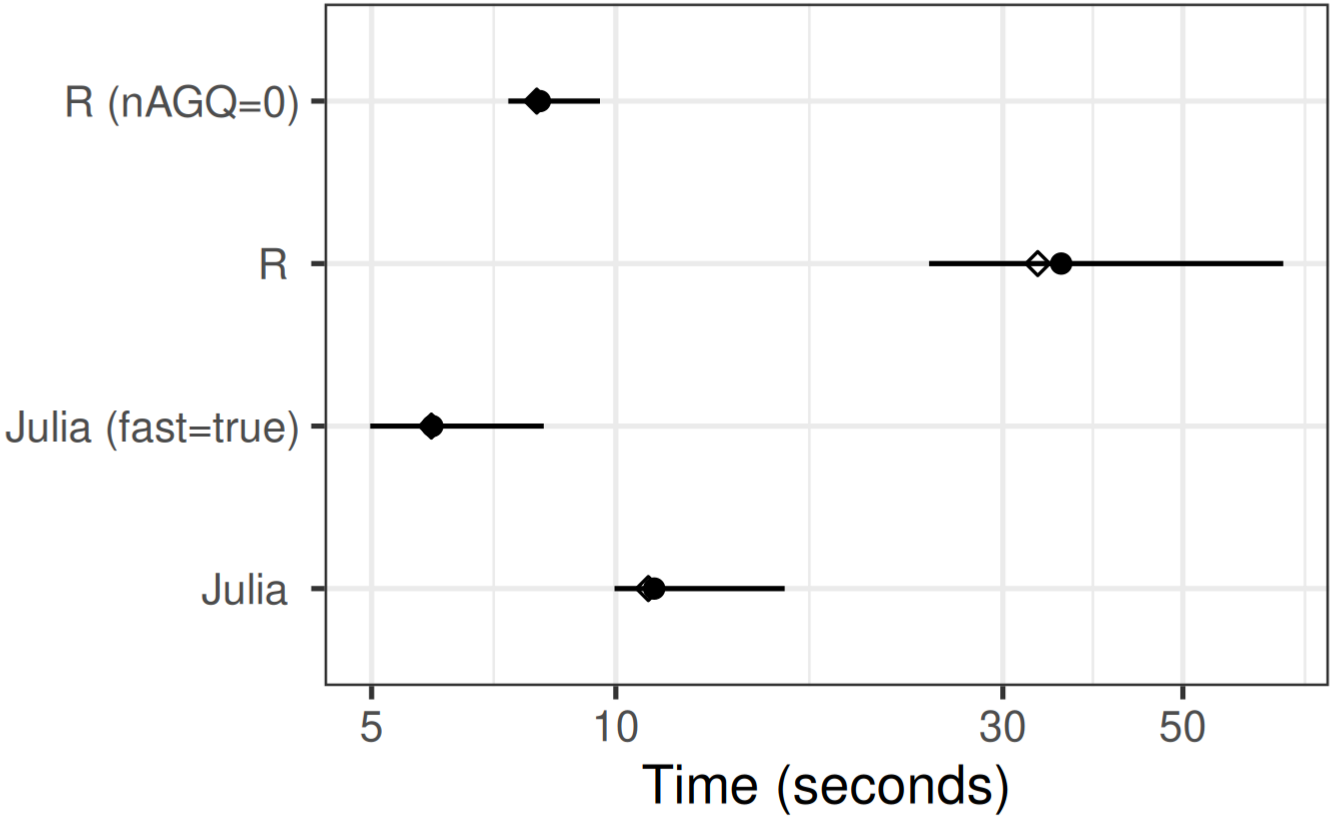


more speed


10000
5000
1000
500
100
50
10
5
1
estimation time (seconds)
23
92
368
1472
5888
11776
random effect levels (sites)
lme4: 0.9
glmmADMB: 1.3
glmmTMB: 1.1
INLA: 1.2
gamlss: 0.7
mgcv: 2.5

Julia/R benchmark

- Poisson model, 98K rows/151 groups (Markwick 2022)



Scalable mixed models

- Gao et al. (2020); Ghosh et al. (2022); Bellio et al. (2025)
- $\mathcal{O}(N)$ methods (`lme4` etc. are $\mathcal{O}(N^{3/2})$)
- StitchFix data: 5M observations, 6.3K items, 762K clients

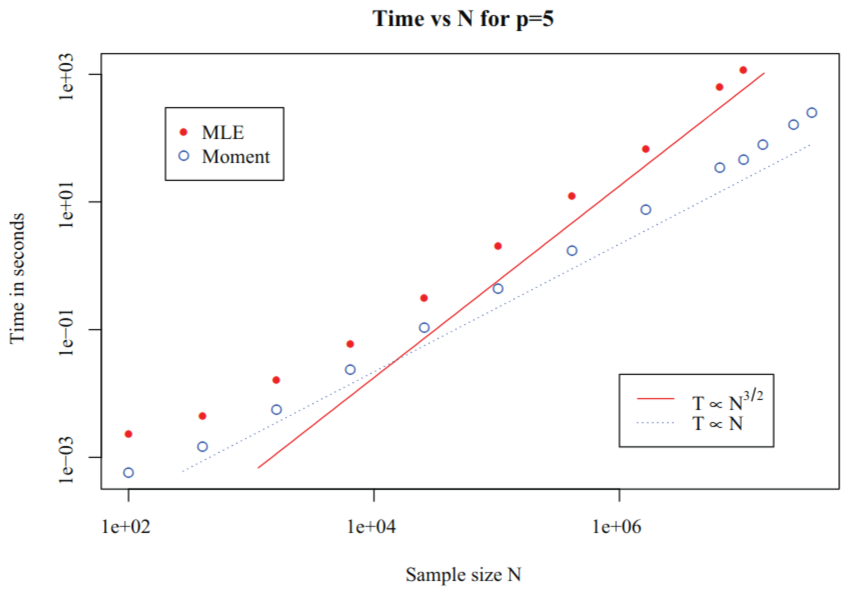


Figure 3. Computational cost for MLE and moments versus sample size N. There are reference lines parallel to $N^{3/2}$ and N^1.

Challenges and open questions

On beyond `lme4`

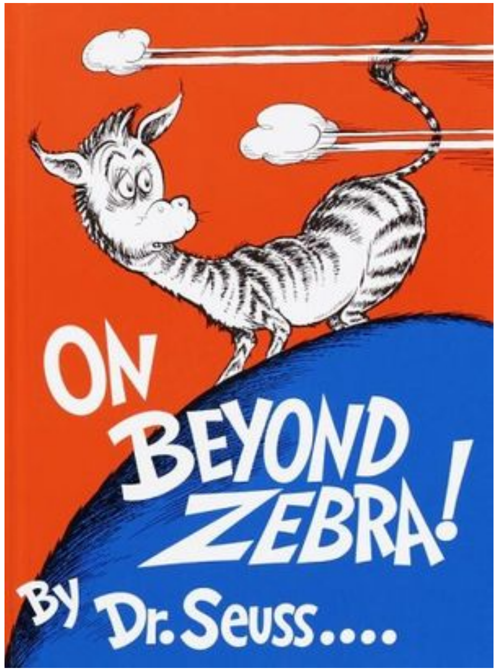
- `glmmTMB`: zero-inflation, weird conditional distributions
- `MCMCglmm`, `brms`,`rstanarm`: Bayesian
- `gllvm`: multivariate ecological/phylogenetic data
- `INLA`: great for spatial and temporal structures
- rethinking package

Toolboxes

- JAGS (R: `rjags`, `r2jags`)
- TensorFlow (R: `greta`)
- NIMBLE (R: `nimble` package)
- Stan (R: `rstan`)
- TMB (R: `TMB`, `RTMB`)

On beyond R

- Julia: `MixedModels.jl`
- SAS: PROC MIXED, NLMIXED
- AS-Reml
- Stata (GLLAMM, xtmelogit)
- HLM, MLWiN
- Python? (`pymer4`, `statsmodels`)



Package comparison

Capability	lme4	GLMMadaptive	glmmTMB	mgcv	nlme	glmmrBase
Linear mixed models	✅	❌	✅	✅	✅	✅
GLMMs	✅	✅	✅	⚠️	❌	✅
Laplace approximation	✅	❌	✅	✅ (PQL)	✅	✅
Gauss–Hermite quadrature	✅ (limited)	✅	❌	❌	❌	❌
Structured covariance (CS/AR1/etc.)	⚠️	❌	✅	✅	✅	✅
Space/time covariance	❌	❌	✅	⚠️	✅	✅
Heteroscedasticity	❌	⚠️	✅	⚠️	✅	✅

Challenges

- Small clusters: need AGQ/MCMC
- Small numbers of clusters: need finite-size corrections
- Small data sets: singular fits, model selection
- Ever-expanding (desired) feature matrix; downstream ecosystems
- Big data: speed!

references

Barr, DJ et al. 2013. *Journal of Memory and Language* 68 (3) (April): 255–278. doi:10.1016/j.jml.2012.11.001.

Bates, D et al. 2015. *arXiv:1506.04967 [stat]* (June). http://arxiv.org/abs/1506.04967.

Bellio, R et al. 2025. *Biometrika* 112 (3) (August): asaf037. doi:10.1093/biomet/asaf037.

Brooks, ME et al. 2017. *R Journal* 9: 378–400. doi:10.32614/RJ-2017-066.

Chung, Y et al. 2013. *Psychometrika*: 1–25. doi:10.1007/s11336-013-9328-2. http://link.springer.com/article/10.1007/s11336-013-9328-2.

Gao, K et al. 2020. *Statistica Sinica*. doi:10.5705/ss.202018.0029.

Ghosh, S et al. 2022. *Electronic Journal of Statistics* 16 (2) (January): 4604–4635. doi:10.1214/22-EJS2047.

Hefley, TJ et al. 2017. *Ecology* 98 (3): 632–646. doi:10.1002/ecy.1674.

Hodges, JS. 2016. *Richly parameterized linear models: Additive, time series, and spatial models using random effects*. Chapman; Hall/CRC.

Markwick, D. 2022. https://dm13450.github.io/2022/01/06/Mixed-Models-Benchmarking.html.

Matuschek, H et al. 2017. *Journal of Memory and Language* 94: 305–315. doi:10.1016/j.jml.2017.01.001.

Moritz, MA et al. 2023. *Ecology Letters* 26 (4): 563–574. doi:10.1111/ele.14177.

Pedersen, EJ et al. 2018. *Hierarchical generalized additive models: An introduction with mgcv*. PeerJ Inc. doi:10.7287/peerj.preprints.27320v1.

Scandola, M et al. 2024. *Advances in Methods and Practices in Psychological Science* 7 (1) (January): 25152459231214454. doi:10.1177/25152459231214454.

Thai, H-T et al. 2013. *Pharmaceutical Statistics* 12 (3): 129–140. doi:10.1002/pst.1561.

Vaida, F et al. 2005. *Biometrika* 92 (2) (June): 351–370. doi:10.1093/biomet/92.2.351. http://biomet.oxfordjournals.org/cgi/content/abstract/92/2/351.

Wahba, G. 1990. *Spline models for observational data*. Philadelphia, Pa.: Society for Industrial; Applied Mathematics.