

# **mixed models: challenges and frontiers**

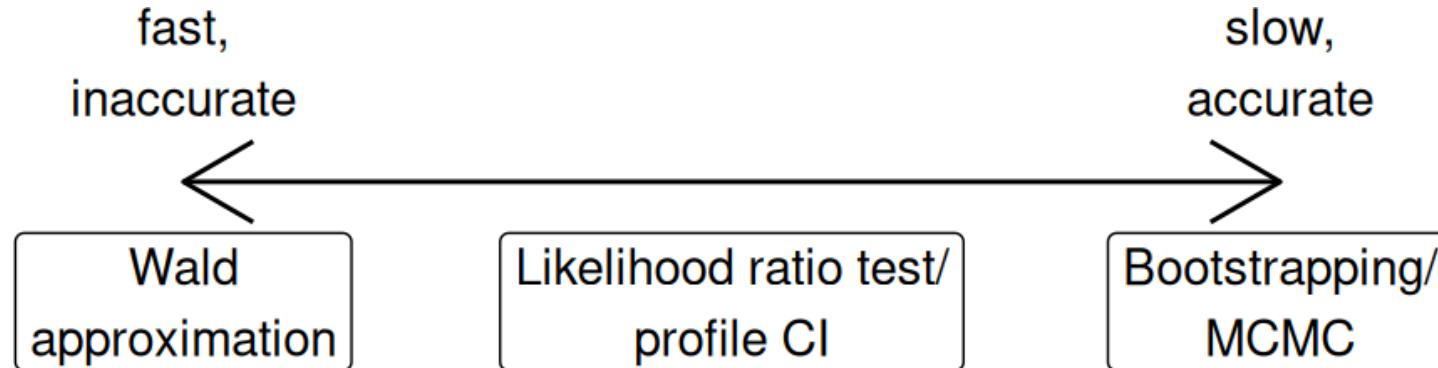
Ben Bolker  
18 November 2025

# Table of contents

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

# Inference details

# Confidence intervals/ $p$ -values



- easy methods work better as number of observations *and* number of clusters grow
- easy methods 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
```

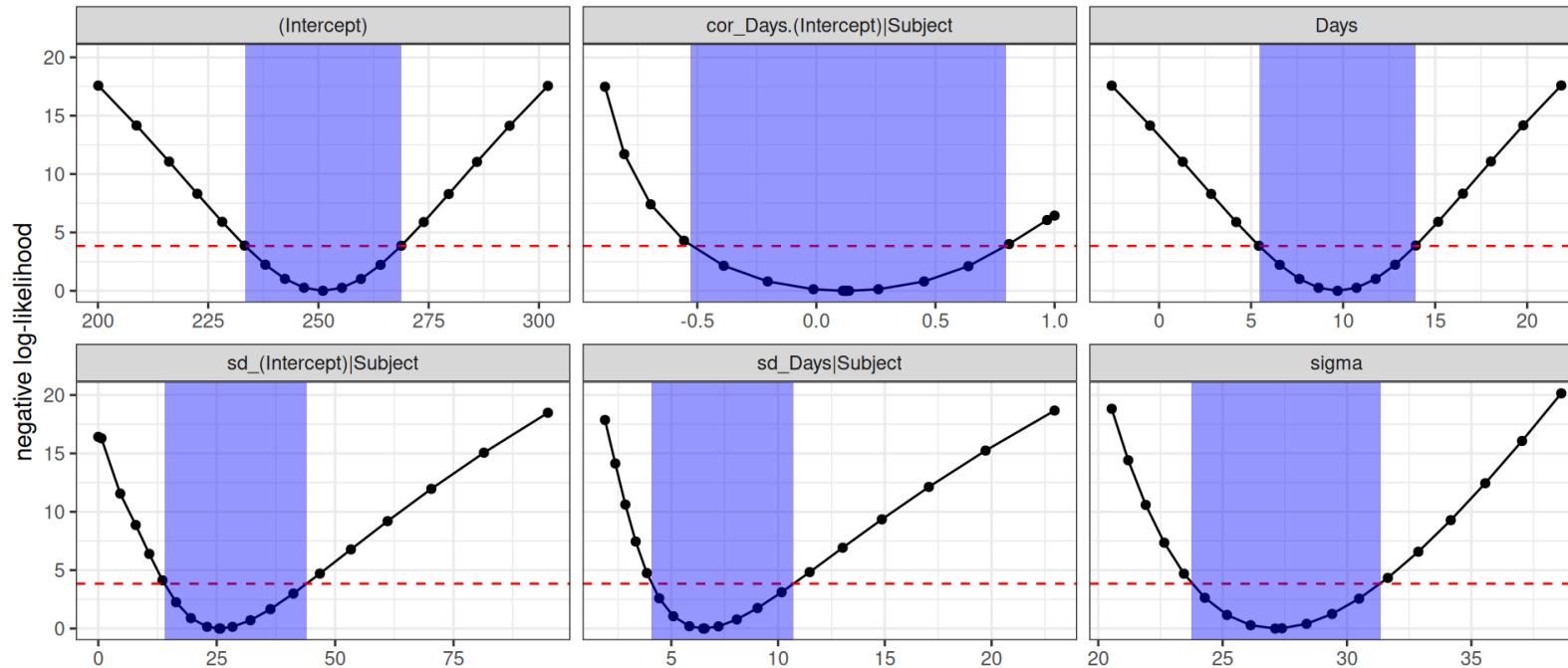
# Likelihood ratio test/profile confidence intervals

- 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

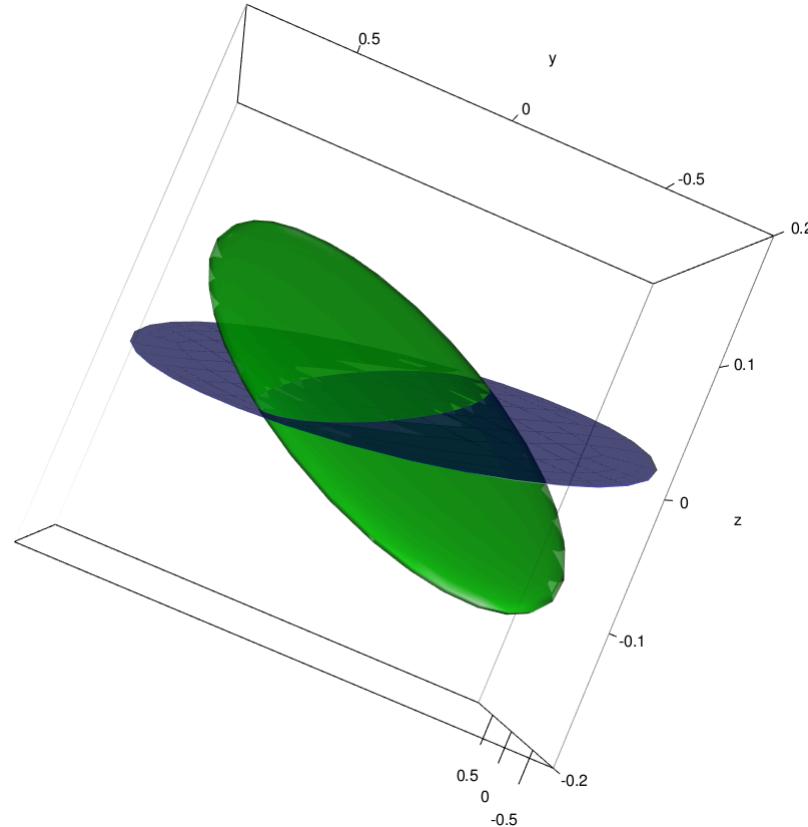
# Predictive simulation

- `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

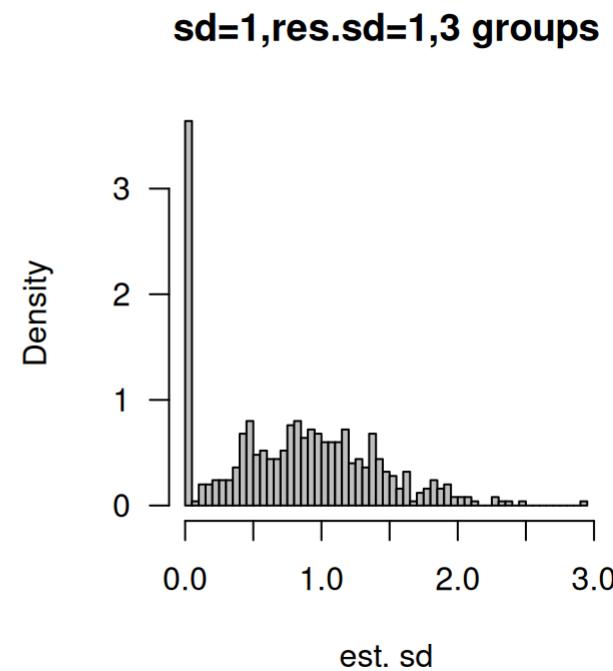
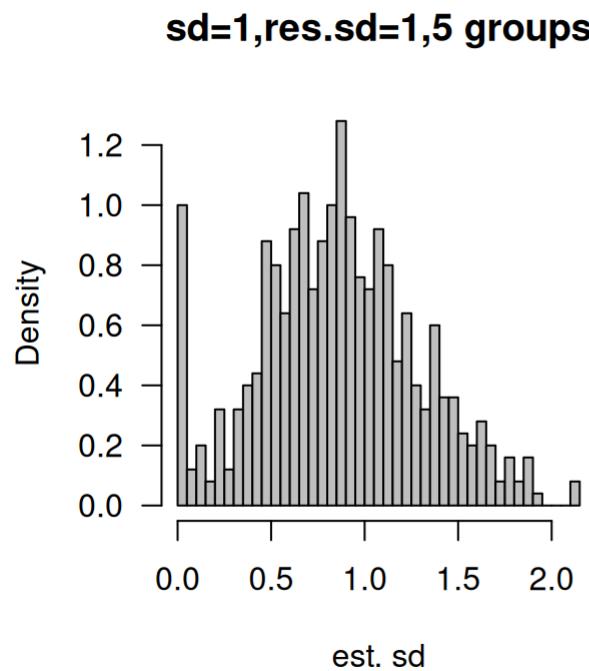
# What is a “singular fit”?

- non-positive-definite covariance matrix
- zero variance (simple case)
- $\pm 1$  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 ( $= \frac{\sigma_{\text{among}}^2 + \sigma_{\text{within}}^2}{n}$ ). 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  $\rightarrow \infty$ )
  - more conservative than dropping/pooling (RE variance = 0)
- e.g. *Arabidopsis*: region (3) / population (9) / genotype (24), use  $\sim \text{reg} + (1 | (\text{reg} : \text{population}) / \text{genotype})$

# 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` + prior specification
- regularization useful in other cases (e.g. 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

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

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

# Sum-to-zero contrasts

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

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

# Complex random intercepts (Scandola et al. 2024)

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

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

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

# Diagonal covariance matrices

- $(1+x+y+z \mid \mid b)$  or `diag(1+x+y+z \mid b)`
  - or  $(1 \mid b) + (0+x \mid b) + \dots$
- `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
- $\Sigma = \Lambda D \Lambda^\top$  ( $\Lambda$  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  $\rho, \rho^2, \rho^3 \dots$  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  $\sigma^2 bSb^\top$
- 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

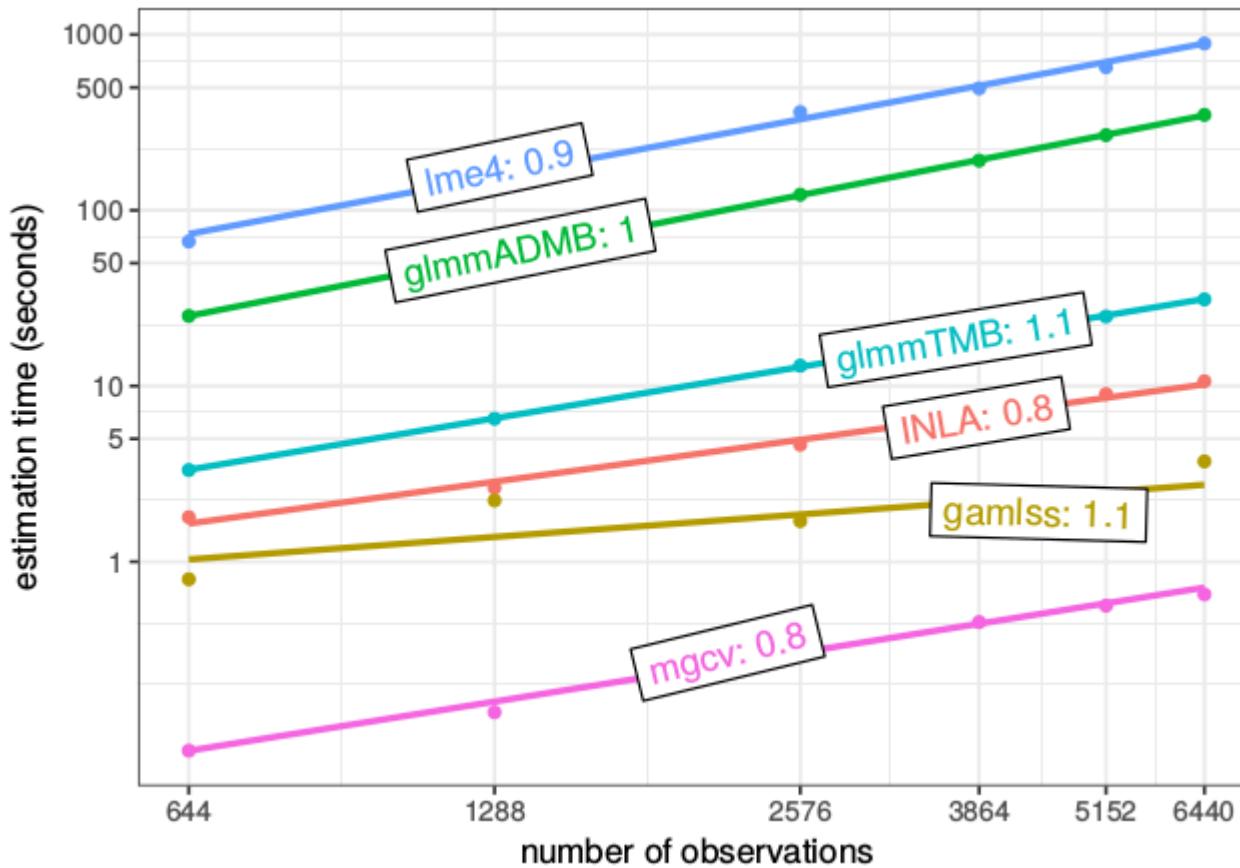
wait, it's all just quadratic  
penalty matrices?



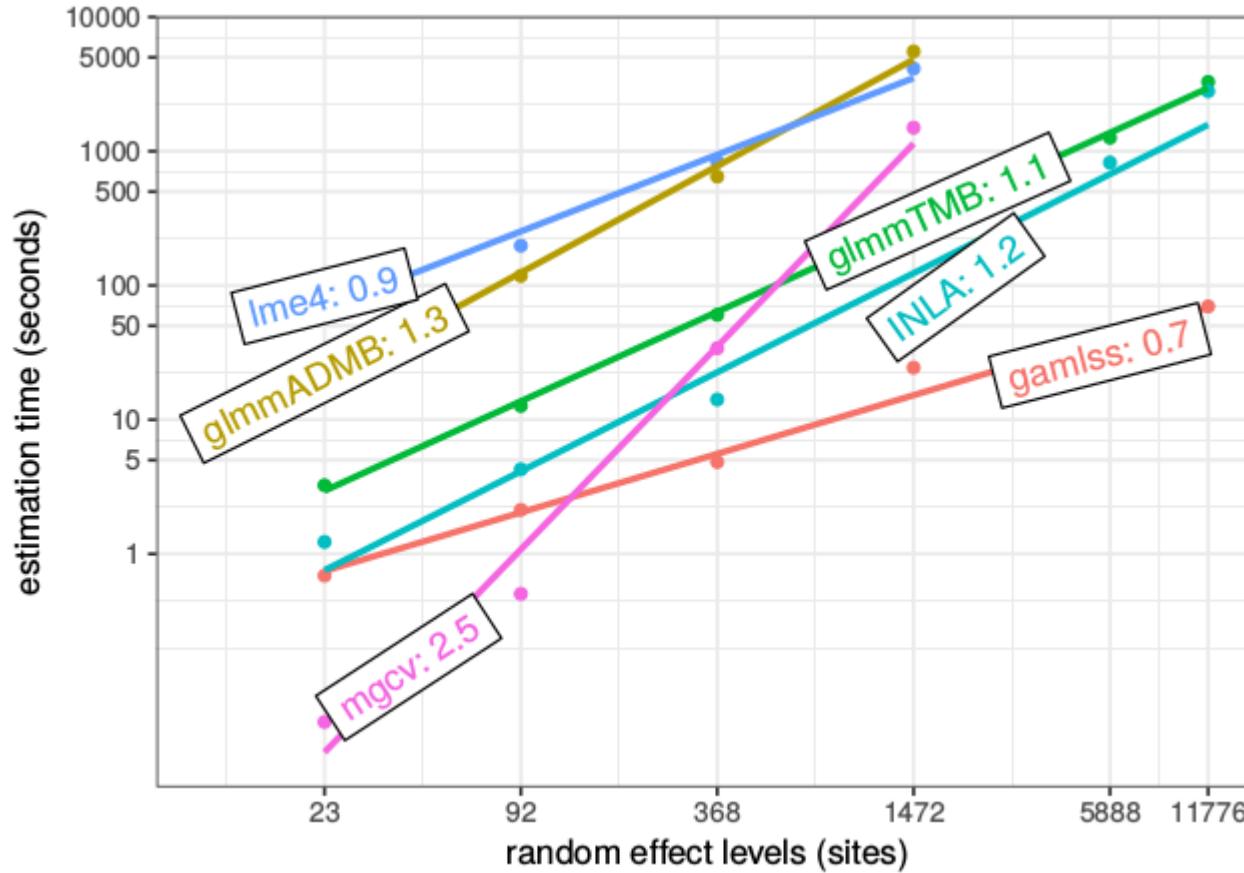
Big data?

# The need for speed

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

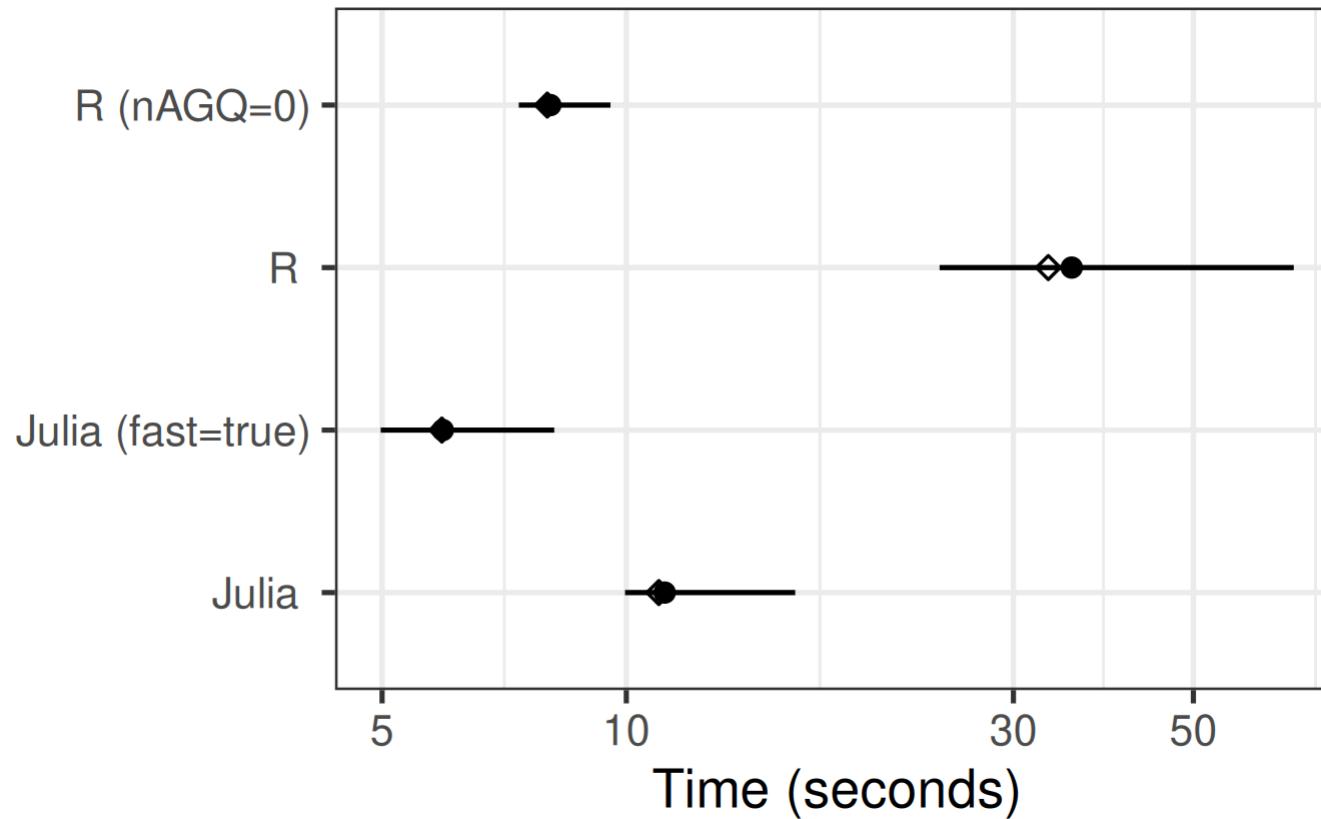


# more speed



# 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

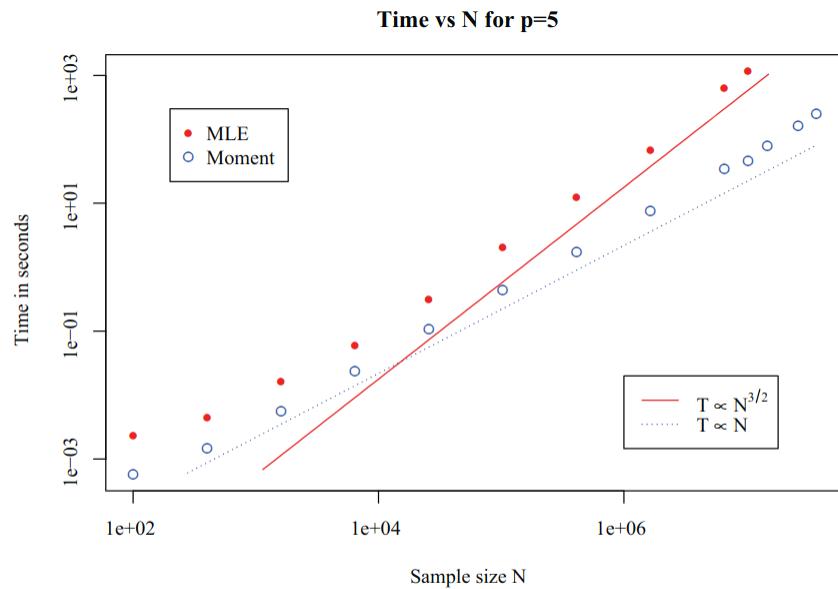


Figure 3. Computational cost for MLE and moments versus sample size  $N$ . There are reference lines parallel to  $N^{3/2}$  and  $N^1$ .

# Challenges and open questions

# On beyond `lme4`

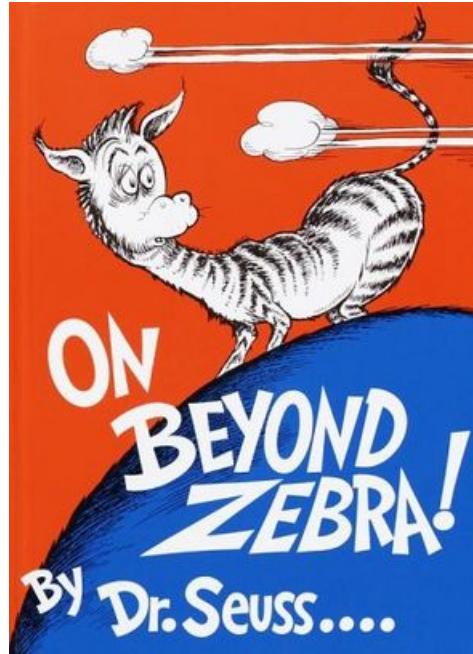
- `glmmTMB`: zero-inflation, weird conditional distributions
- `MCMCglmm`, `brms`, `rstanarm`: Bayesian
- `gllvm`: multivariate ecological/phylogenetic data
- `INLA`: great for spatial and temporal structures
- [rethinking package](#)

# 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](https://doi.org/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](https://doi.org/10.1093/biomet/asaf037).
- Brooks, ME et al. 2017. *R Journal* 9: 378–400. doi:[10.32614/RJ-2017-066](https://doi.org/10.32614/RJ-2017-066).
- Byrnes, JEK et al. 2025. *Ecology Letters* 28 (1): e70023. doi:[10.1111/ele.70023](https://doi.org/10.1111/ele.70023).  
<https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.70023>.
- Chung, Y et al. 2013. *Psychometrika*: 1–25. doi:[10.1007/s11336-013-9328-2](https://doi.org/10.1007/s11336-013-9328-2).
- Elston, DA et al. 2001. *Parasitology* 122 (5): 563–569.
- Franks, DW et al. 2025. *Biological Reviews* 100 (5): 1950–1969. doi:[10.1111/brv.70029](https://doi.org/10.1111/brv.70029).  
<https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.70029>.
- Gao, K et al. 2020. *Statistica Sinica*. doi:[10.5705/ss.202018.0029](https://doi.org/10.5705/ss.202018.0029).
- Ghosh, S et al. 2022. *Electronic Journal of Statistics* 16 (2) (January): 4604–4635. doi:[10.1214/22-EJS2047](https://doi.org/10.1214/22-EJS2047).
- Harrison, XA. 2014. *PeerJ* 2 (October): e616. doi:[10.7717/peerj.616](https://doi.org/10.7717/peerj.616). <https://peerj.com/articles/616>.
- . 2015. *PeerJ* 3 (July): e1114. doi:[10.7717/peerj.1114](https://doi.org/10.7717/peerj.1114). <https://peerj.com/articles/1114>.
- He, H et al. 2019. *Statistical methods in medical research* 28 (4) (April): 1157–1169. doi:[10.1177/0962280217749991](https://doi.org/10.1177/0962280217749991).
- Hefley, TJ et al. 2017. *Ecology* 98 (3): 632–646. doi:[10.1002/ecy.1674](https://doi.org/10.1002/ecy.1674).
- Hilbe, Joseph M., ed. 2011. In *Negative Binomial Regression*, 141–184. 2nd ed. Cambridge: Cambridge University Press.  
doi:[10.1017/CBO9780511973420.008](https://doi.org/10.1017/CBO9780511973420.008). <https://www.cambridge.org/core/books/negative-binomial-regression/overdispersion/1779F5E5E1CBA5B75B1410BE6ECCA3B>.
- Hodges, JS. 2016. *Richly parameterized linear models: Additive, time series, and spatial models using random effects*. Chapman; Hall/CRC.

**extras**

# $R^2$

- goodness-of-fit measures for LMMs, and for GLMMs, are hard
- meanings of  $R^2$  (variance explained;  $\text{corr}(\text{obs}, \text{predicted})^2$ ; etc. don't coincide)
- *pseudo-R<sup>2</sup>* for GLMs (MacFadden, Nagelkerke, etc.)
- various [Jaeger et al. (2017)]
- Nakagawa/Schielzeth/Johnson most popular (Nakagawa et al. 2013; Johnson 2014; Nakagawa et al. 2017)
- (?but what about zero-inflation, new distributions, etc. etc. etc.)

# Overdispersion

- “too much” variability (relative to theoretical expectation)
- very serious for fixed-dispersion families (Poisson, binomial)
- conjugate-family distributions (negative binomial, Beta-binomial)
- observation-level random effects (Elston et al. 2001; Harrison 2014; Harrison 2015)
- unidentifiable for binary data (unless groupable), Gaussian data
- little concern for families with estimated scale/dispersion (but see Hilbe (2011))

# Zero-inflation

- “many zeros does not mean zero-inflation” (Warton 2005)
- can’t use Vuong test (Wilson 2015; He et al. 2019)
- `glmmTMB` or use expectation-maximization (see [here](#))
- `DHARMA` can test by simulation

# Causal inference

- I don't go there very much
- Solve many issues by adding *group mean* as a covariate (Byrnes et al. 2025)
- Awareness of *general* causal issues: DAGs, mediation/confounding, etc.. (McElreath 2015; Franks et al. 2025)

# Heteroscedasticity

- can be an issue (especially among groups)
- log transformation (for continuous positive data) often resolves both linearity and heteroscedasticity
- `glmmTMB`, `nlme`
- build into observation-level REs