Statistics, computation, and software engineering: development and maintenance of mixed modeling software in R

Ben Bolker

McMaster University, Mathematics & Statistics and Biology

15 October 2013
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
(Generalized) linear mixed models

(G)LMMs: a statistical modeling framework incorporating:

- **Linear combinations** of categorical and continuous predictors, and interactions
- Response distributions in the *exponential family* (binomial, Poisson, and extensions)
- Any smooth, monotonic **link function** (e.g. logistic, exponential models)
- Flexible combinations of **blocking factors** (clustering; random effects)

Applications in ecology, neurobiology, behaviour, epidemiology, real estate, ...
(Generalized) linear mixed models

(G)LMMs: a statistical modeling framework incorporating:

- **Linear combinations** of categorical and continuous predictors, and interactions
- Response distributions in the **exponential family** (binomial, Poisson, and extensions)
- Any smooth, monotonic **link function** (e.g. logistic, exponential models)
- Flexible combinations of **blocking factors** (clustering; random effects)

Applications in ecology, neurobiology, behaviour, epidemiology, real estate, ...
(Generalized) linear mixed models

(GL)MMs: a statistical modeling framework incorporating:

- **Linear combinations** of categorical and continuous predictors, and interactions
- Response distributions in the **exponential family**
  (binomial, Poisson, and extensions)
- Any smooth, monotonic **link function**
  (e.g. logistic, exponential models)
- Flexible combinations of **blocking factors**
  (clustering; random effects)

Applications in ecology, neurobiology, behaviour, epidemiology, real estate, ...
(Generalized) linear mixed models

(G)LMMs: a statistical modeling framework incorporating:

- **Linear combinations** of categorical and continuous predictors, and interactions
- Response distributions in the **exponential family** (binomial, Poisson, and extensions)
- Any smooth, monotonic **link function** (e.g. logistic, exponential models)
- Flexible combinations of **blocking factors** (clustering; random effects)

Applications in ecology, neurobiology, behaviour, epidemiology, real estate, ...
Examples

- **ecology**  survival, predation, etc. (experimental plots)
- **genomics**  presence/absence of polymorphisms, gene expression (individuals)
- **educational assessment**  student scores (students × teachers)
- **psychology/sensometrics**  decisions, responses to stimuli (individuals)
- **epidemiology**  disease prevalence (postal codes, provinces, countries)
Examples

- **ecology**  survival, predation, etc. (experimental plots)
- **genomics** presence/absence of polymorphisms, gene expression (individuals)
- educational assessment  student scores (students × teachers)
- psychology/sensometrics  decisions, responses to stimuli (individuals)
- epidemiology  disease prevalence (postal codes, provinces, countries)
Examples

- **ecology** survival, predation, etc. (experimental plots)
- **genomics** presence/absence of polymorphisms, gene expression (individuals)
- **educational assessment** student scores (students × teachers)
- **psychology/sensometrics** decisions, responses to stimuli (individuals)
- **epidemiology** disease prevalence (postal codes, provinces, countries)
Examples

- **ecology** survival, predation, etc. (experimental plots)
- **genomics** presence/absence of polymorphisms, gene expression (individuals)
- **educational assessment** student scores (students × teachers)
- **psychology/sensometrics** decisions, responses to stimuli (individuals)
- **epidemiology** disease prevalence (postal codes, provinces, countries)
Examples

ecology  survival, predation, etc. (experimental plots)
genomics  presence/absence of polymorphisms, gene expression (individuals)
educational assessment  student scores (students $\times$ teachers)
psychology/sensometrics  decisions, responses to stimuli (individuals)
epidemiology  disease prevalence (postal codes, provinces, countries)
Technical definition

\[ Y_i \sim \text{Distr} \left( g^{-1}(\eta_i), \phi \right) \]

\[ \eta = X\beta + Zb \]

\[ b \sim \text{MVN}(0, \Sigma(\theta)) \]
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
Estimation

Maximum likelihood estimation

\[
\mathcal{L}(Y_i|\theta, \beta) = \int \ldots \int \mathcal{L}(Y_i|\theta, \beta') \times \mathcal{L}(\beta'|\Sigma(\theta)) \, d\beta'
\]

- **Monte Carlo**: frequentist and Bayesian (Booth and Hobert, 1999; Ponciano et al., 2009; Sung, 2007)
Estimation: example (McKeon et al., 2012)

Log-odds of predation

Added symbiont
Crab vs. Shrimp
Symbiont

GLM (fixed)
GLM (pooled)
PQL
Laplace
AGQ

Ben Bolker
Mixed model software
Inference

- **Standard inferential tools:** mostly asymptotic or uncontrolled approximations
- **Solutions** are computational and/or Bayesian: parametric bootstrap, MCMC
- **Good news:** different problems for small vs large data
Inference

- **Standard inferential tools:** mostly asymptotic or uncontrolled approximations
- **Solutions are computational** and/or Bayesian: parametric bootstrap, MCMC
- **Good news:** different problems for small vs large data
Inference

- Standard inferential tools: mostly asymptotic or uncontrolled approximations
- Solutions are computational and/or Bayesian: parametric bootstrap, MCMC
- Good news: different problems for small vs large data
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
Problems of big data

- How big is big?
  Airline data: 12G

- (G)LMM works on moderately large problems,
  e.g. student evaluations
  (∼ 75K total, 3K students, 1K profs)

- Fairly clever linear algebra

- Possible improvements?
  - Chunking/parallelization
  - Out-of-memory operation
Sparse matrix algorithms

- repeated decomposition of large, matrices (especially $Z$)
- fill-reducing permutation to improve sparsity pattern
- further improvements possible: better matrix representation, parallelization?
Bounded optimization

- Parameterize variance-covariance matrix $\Sigma(\theta)$ (Pinheiro and Bates, 1996)
- Positive definite or only semi-definite?
- Disadvantages of transforming to unconstrain
- (Disadvantages of boundary solutions)
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
Language tradeoffs

- high-level/convenience: R
- low-level/performance: C++
- new wave? Julia
- multi-language friction: mostly escaped in R/C++ case, at the price of complexity
Getting it right vs. getting it written

- the curse of neophilia: **Superiority**
- many versions: `nlme`, `lme4(a,b,Eigen)` . . .
- *The moral of the story is that if you want to create a beautiful language, for god’s sake don’t make it useful*  
  *(Patrick Burns)*
Sociological issues

- **Wide user base:**
  
  *As usual when software for complicated statistical inference procedures is broadly disseminated, there is potential for abuse and misinterpretation.*
  
  *(Breslow, 2004)*

- **What if there is no good answer?**
  
  “do no harm” vs. “better me than someone else”

- **Diagnostics and warning messages**

- **End users vs. downstream developers**
Outline

1. Definitions and context
2. Statistical challenges
3. Computational challenges
4. Software engineering
5. Conclusions
Next steps

- Alternative platforms/languages
- Flexible correlation structures: spatial, temporal, phylogenetic . . .
- Improved MCMC methods?
- Simulation tests of inferential tools
Is it science?

Science is what we understand well enough to explain to a computer. Art is everything else we do. (Donald Knuth)
Acknowledgments

- **lme4**: Doug Bates, Martin Mächler, Steve Walker
- **Data**: Adrian Stier (UBC/OSU), Sea McKeon (Smithsonian), David Julian (UF)
- **NSERC (Discovery)**
- **SHARCnet**


