generalized linear mixed models

Ben Bolker

(Generalized) linear mixed models

(G)LMMs: a statistical modeling framework incorporating:

- combinations of categorical and continuous predictors, and interactions
- (some) non-Normal responses (e.g. binomial, Poisson, and extensions)
- (some) nonlinearity (e.g. logistic, exponential, hyperbolic)
- non-independent (grouped) data

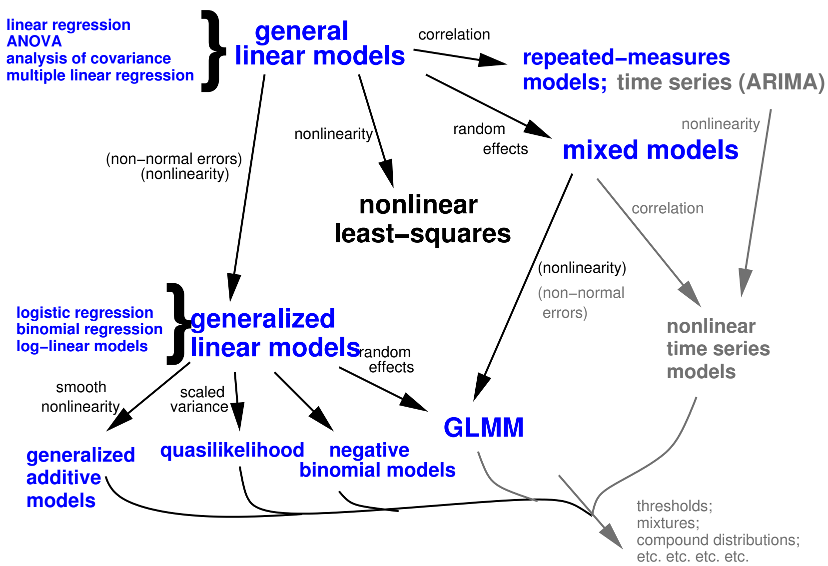
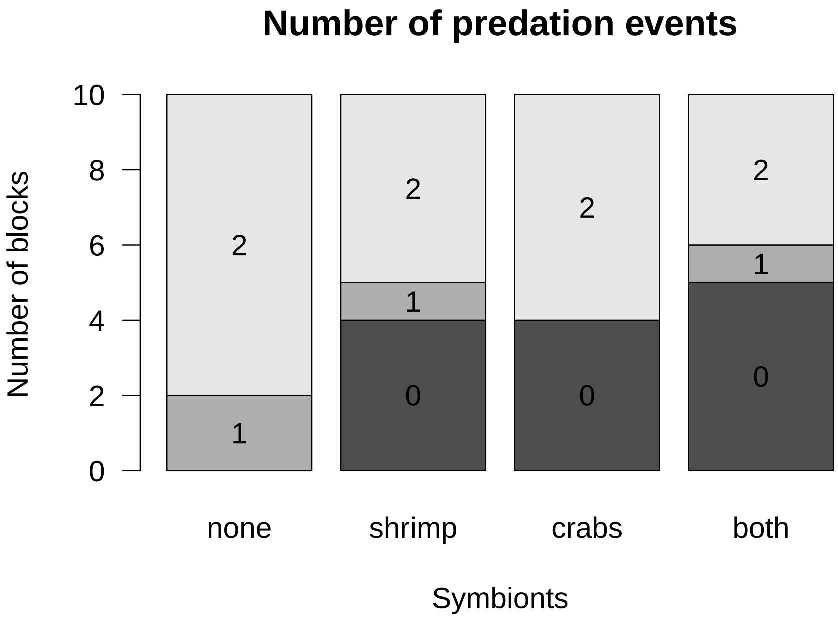
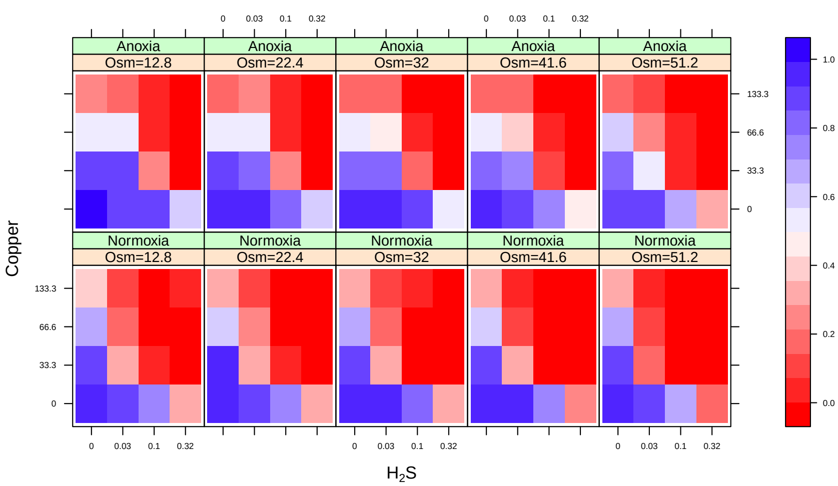


Figure 1: image

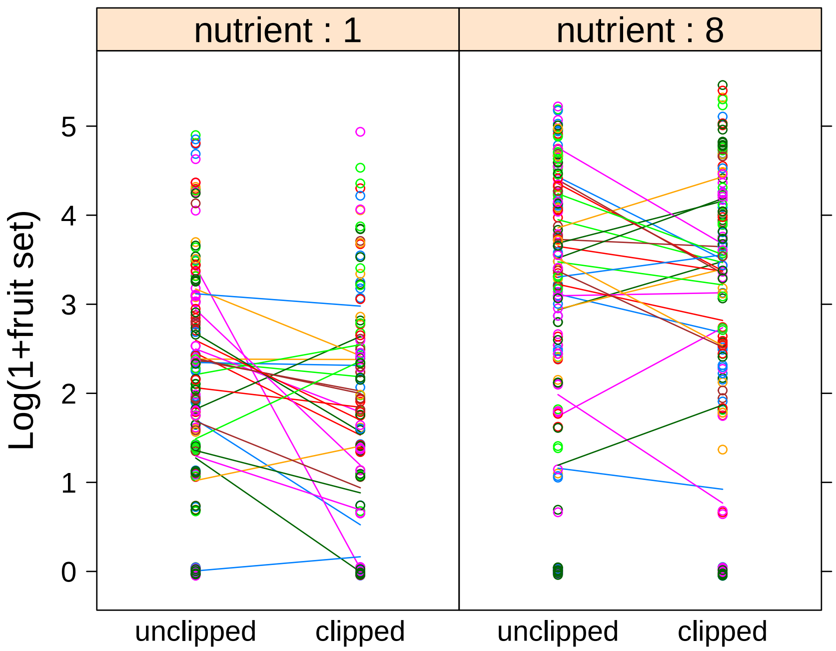
Coral protection from seastars (Culcita) *by symbionts (McKeon et al. 2012)*



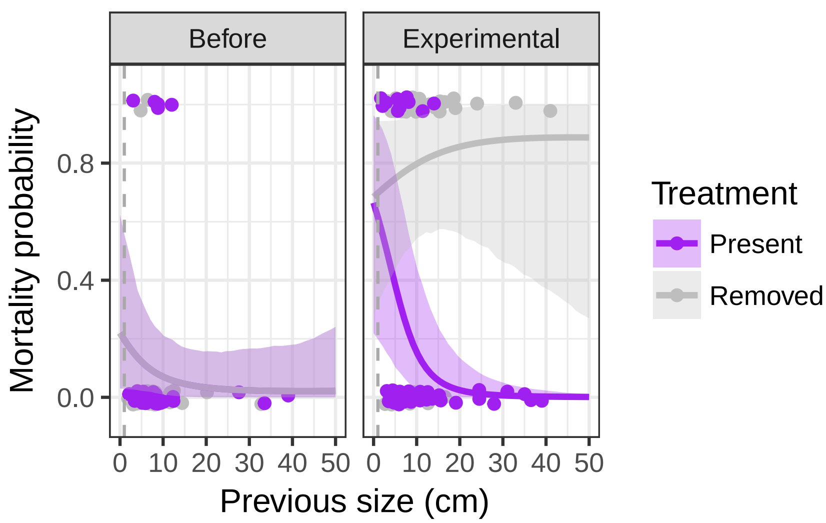
Environmental stress: Glycera *cell survival (D. Julian unpubl.)*



Arabidopsis *response to fertilization & herbivory (Banta, Stevens, and Pigliucci 2010)*



Coral demography (J.-S. White unpubl.)



Technical definition

$$
\begin{aligned}
\underbrace{Y_i}_{\text{response}} &\sim \overbrace{\text{Distr}}^{\substack{\text{conditional}\\\text{distribution}}} \left(\underbrace{g^{-1}(\eta_i)}_{\substack{\text{inverse}\\\text{link}\\\text{function}}}, \underbrace{\phi}_{\substack{\text{scale}\\\text{parameter}}}\right) \\
\underbrace{\boldsymbol{\eta}}_{\substack{\text{linear}\\\text{predictor}}} &= \underbrace{\boldsymbol{X\beta}}_{\substack{\text{fixed}\\\text{effects}}} + \underbrace{\boldsymbol{Zb}}_{\substack{\text{random}\\\text{effects}}} \\
\underbrace{\boldsymbol{b}}_{\substack{\text{conditional}\\\text{modes}}} &\sim \text{MVN}\left(\boldsymbol{0}, \underbrace{\Sigma(\boldsymbol{\theta})}_{\substack{\text{variance-}\\\text{covariance}\\\text{matrix}}}\right)
\end{aligned}
$$

What are random effects?

A method for ...

- accounting for among-individual, within-block correlation
- compromising between
  complete pooling (no among-block variance)
  and *fixed effects* (large among-block variance)
- handling levels selected at random from a larger population
- sharing information among levels (`*shrinkage estimation*`)
- estimating variability among levels
- allowing predictions for unmeasured levels

Random-effect myths

- levels of random effects must always be sampled at random
- a complete sample cannot be treated as a random effect
- random effects are always a *nuisance variable*
- nothing can be said about the predictions of a random effect
- you should always use a random effect no matter how few levels you have

Use a random effect if:

(from B. M. Bolker (2015))

- don't want to test hypotheses about differences between responses at particular levels of the grouping variable;
- do want to quantify the variability among levels of the grouping variable;

- do want to make predictions about unobserved levels of the grouping variable;
- do want to combine information across levels of the grouping variable;
- have variation in information per level (number of samples or noisiness);
- have levels that are randomly sampled from/representative of a larger population;
- have a categorical predictor that is a nuisance variable (i.e., it is not of direct interest, but should be controlled for).

See also Crawley (2002); Gelman (2005)

If you have sampled fewer than five levels of the grouping variable, you should strongly consider treating it as a fixed effect even if one or more of the criteria above apply.

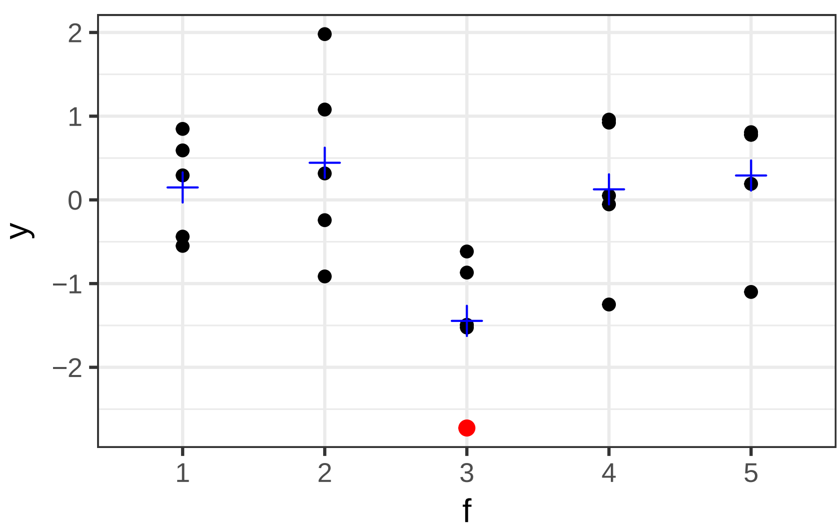
Estimation

Overview

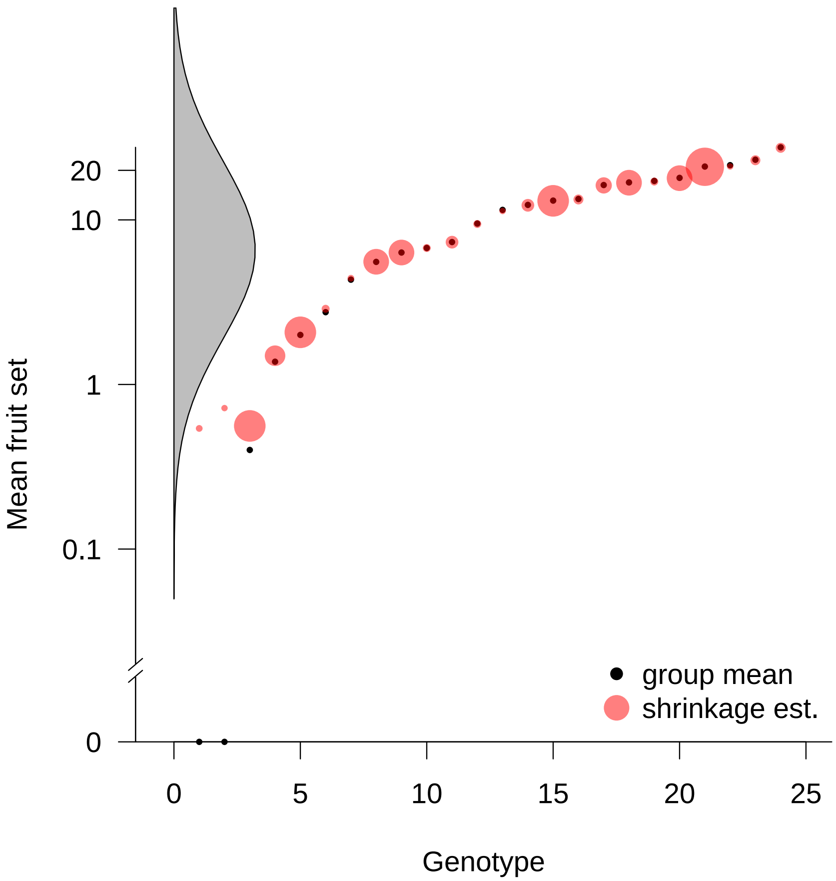
Maximum likelihood estimation

- Best fit is a compromise between two components (consistency of data with fixed effects and conditional modes; consistency of random effect with RE distribution)

- Goodness-of-fit `*integrates*` over conditional modes



Shrinkage: Arabidopsis *conditional modes*

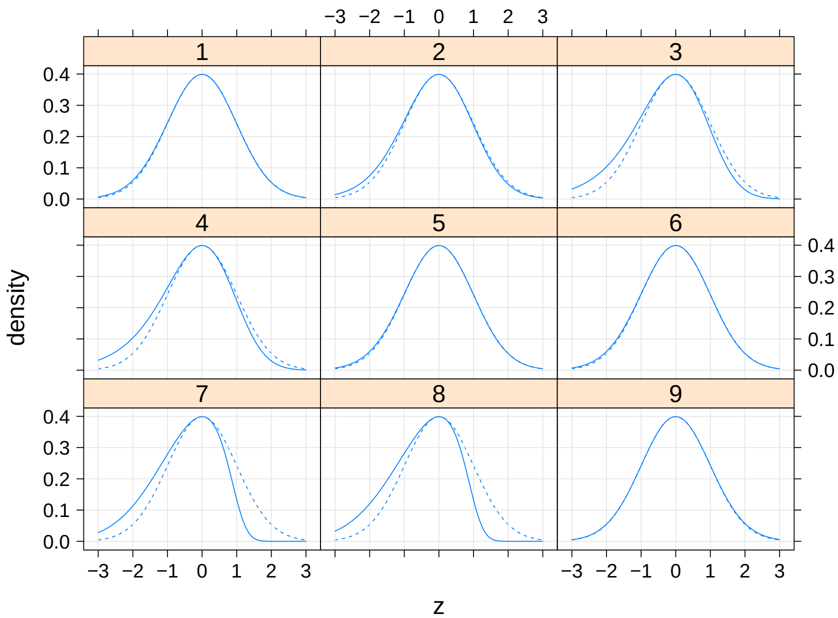


Methods

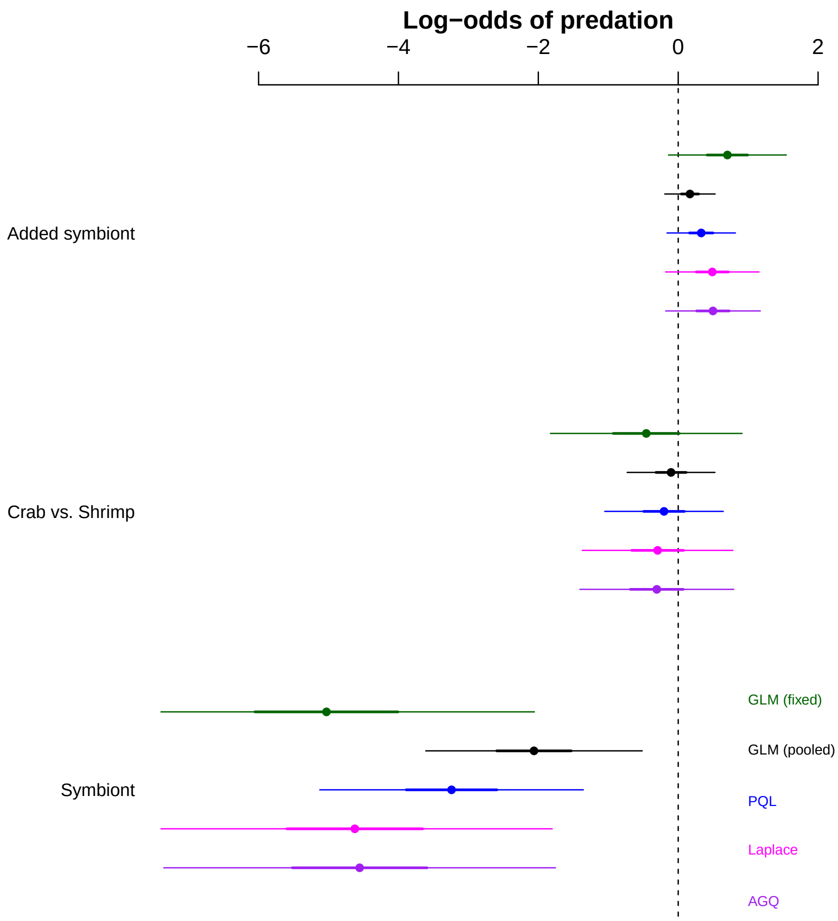
Estimation methods

- deterministic
  - various approximate integrals (Breslow 2004)
  - penalized quasi-likelihood, Laplace, Gauss-Hermite quadrature, . . . (Biswas 2015);
    best methods needed for large variance, small clusters
  - flexibility and speed vs. accuracy
- stochastic
- stochastic (Monte Carlo): frequentist and Bayesian
  - (Booth and Hobert 1999; Sung and Geyer 2007; Ponciano et al. 2009)
  - usually slower but flexible and accurate

Laplace-approximation diagnostics



Estimation: Culcita *(McKeon et al. 2012)*



Inference

Wald tests

- typical results of `summary`
- exact for ANOVA, regression: approximation for GLM(M)s
- fast
- approximation is sometimes awful (Hauck-Donner effect)

Likelihood ratio tests

- better than Wald, but still have two problems:
  - "denominator degrees of freedom" (when estimating scale)
  - for GLMMs, distributions are approximate anyway (Bartlett corrections)
  - Kenward-Roger correction? (Stroup 2014)

- Profile confidence intervals: expensive/fragile

p-values choices?

- guess from classic design (R code)
- conservative: take minimum number of groups - 1
- Satterthwaite/Kenward-Roger (`lmerTest`, LMMs only)
- parametric bootstrap (`pbkrtest`)

Parametric bootstrapping

- fit null model to data
- simulate "data" from null model
- fit null and working model, compute likelihood difference
- repeat to estimate null distribution
- should be OK but ??? not well tested
  (assumes estimated parameters are "sufficiently" good)

Bayesian inference

- If we have a good sample from the posterior distribution (Markov chains have converged etc. etc.) we get most of the inferences we want for free by summarizing the marginal posteriors
- `*post hoc*` Bayesian methods: use deterministic/frequentist methods to find the maximum, then sample around it

Culcita *confidence intervals*

formula formats

- `fixed`: fixed-effect formula
- `random`: random-effect formula (in `lme4`, combined with fixed)
  - generally `x|g` (term|grouping variable)
  - simplest: `1|g`, single intercept term
  - nested: `1|g1/g2`
  - random-slopes: `r|g`
  - independent terms: `(1|g)+(x+0|g)` or `(x||g)`
- `lme`: `weights`, `correlation` for heteroscedasticity and residual correlation
- `MCMCglmm`: options for variance structure

Challenges & open questions

On beyond `lme4`

- `glmmTMB`: zero-inflated and other distributions

- `brms`,`rstanarm`: interfaces to Stan
- `INLA`: spatial and temporal correlations

On beyond R

- Julia: MixedModels package
- SAS: PROC MIXED, NLMIXED
- AS-REML
- Stata (GLLAMM, xtmelogit)
- AD Model Builder; Template Model Builder
- HLM, MLWiN
- JAGS, Stan, rethinking package

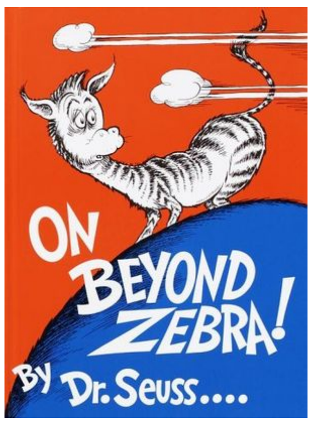


Figure 2: image

Challenges

- Small clusters: need AGQ/MCMC
- Small numbers of clusters: need finite-size corrections (KR/PB/MCMC)
- Small data sets: issues with *singular* fits
  (Barr et al. 2013) vs. (Bates et al. 2015)
- Big data: speed!
- Model diagnosis
- Confidence intervals accounting for uncertainty in variances

See also: `https://rawgit.com/bbolker/mixedmodels-misc/master/ecostats_chap.html` `https://groups.nceas.ucsb.edu/non-linear-modeling/projects`

Spatial and temporal correlations

- Sometimes blocking takes care of non-independence . . .
- but sometimes there is temporal or spatial correlation *within* blocks
- . . . also phylogenetic . . . (Ives and Zhu 2006)

- "G-side" vs. "R-side" effects
- tricky to implement for GLMMs, but new possibilities on the horizon (Rue, Martino, and Chopin 2009; Rousset and Ferdy 2014); https://github.com/stevencarlislewalker/lme4ord

Next steps

- Complex random effects:
  regularization, model selection, penalized methods (lasso/fence)
- Flexible correlation and variance structures
- Flexible/nonparametric random effects distributions
- hybrid & improved MCMC methods
- *Reliable* assessment of out-of-sample performance

- `http://ms.mcmaster.ca/~bolker/misc/private/14-Fox-Chap13.pdf`
- `https://rawgit.com/bbolker/mixedmodels-misc/master/ecostats_chap.html`
- (B. M. Bolker 2015)

  (code ASPROMP8)

Banta, Joshua A., Martin H. H. Stevens, and Massimo Pigliucci. 2010. "A Comprehensive Test of the 'Limiting Resources' Framework Applied to Plant Tolerance to Apical Meristem Damage." *Oikos* 119 (2): 359–69. doi:10.1111/j.1600-0706.2009.17726.x.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *Journal of Memory and Language* 68 (3): 255–78. doi:10.1016/j.jml.2012.11.001.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2015. "Parsimonious Mixed Models." *ArXiv:1506.04967 [Stat]*, June. `http://arxiv.org/abs/1506.04967`.

Biswas, Keya. 2015. "Performances of Different Estimation Methods for Generalized Linear Mixed Models." Master's thesis, McMaster University. `https://macsphere.mcmaster.ca/bitstream/11375/17272/2/M.Sc_Thesis_final_Keya_Biswas.pdf`.

Bolker, Benjamin M. 2015. "Linear and Generalized Linear Mixed Models." In *Ecological Statistics: Contemporary Theory and Application*, edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa. Oxford University Press.

Booth, James G., and James P. Hobert. 1999. "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte

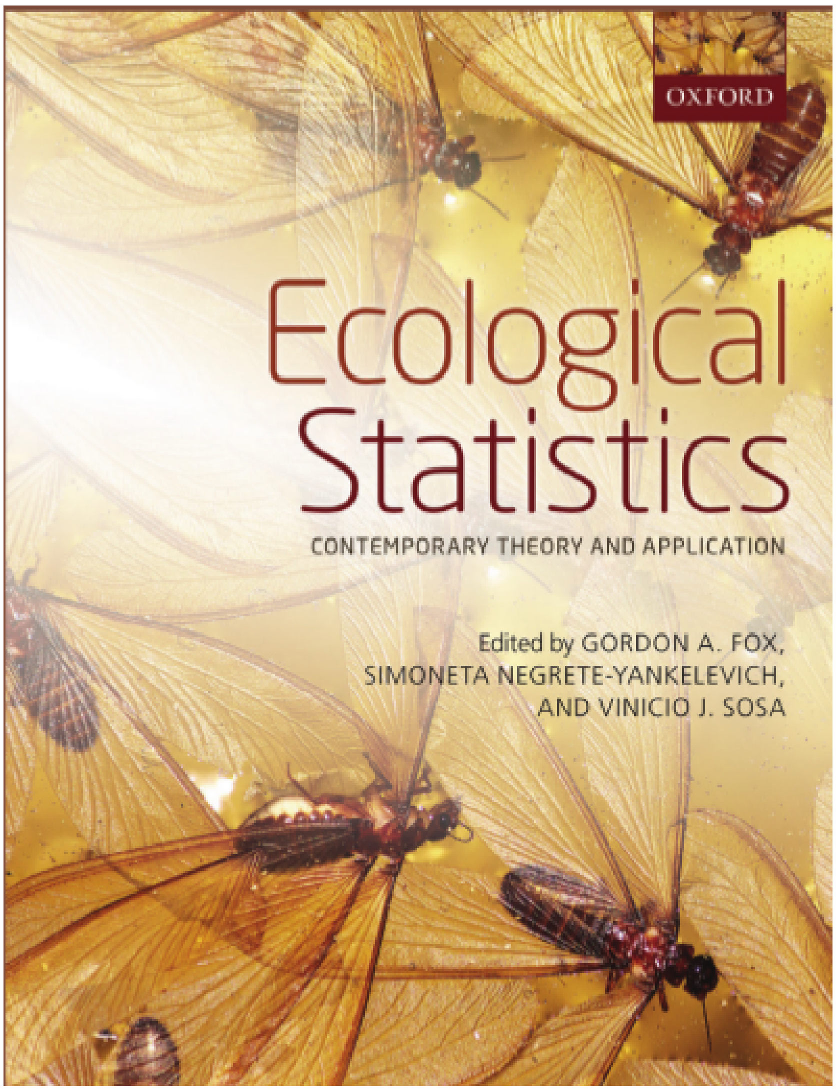


Figure 3: image

Carlo EM Algorithm." *Journal of the Royal Statistical Society. Series B* 61 (1): 265–85. doi:10.1111/1467-9868.00176.

Breslow, N. E. 2004. "Whither PQL?" In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, edited by Danyu Y. Lin and P. J. Heagerty, 1–22. Springer.

Crawley, Michael J. 2002. *Statistical Computing: An Introduction to Data Analysis Using S-PLUS*. John Wiley & Sons.

Gelman, Andrew. 2005. "Analysis of Variance: Why It Is More Important Than Ever." *Annals of Statistics* 33 (1): 1–53. doi:doi:10.1214/009053604000001048.

Ives, Anthony R., and Jun Zhu. 2006. "Statistics for Correlated Data: Phylogenies, Space, and Time." *Ecological Applications* 16 (1): 20–32. `http://www.esajournals.org/doi/pdf/10.1890/04-0702`.

McKeon, C. Seabird, Adrian Stier, Shelby McIlroy, and Benjamin Bolker. 2012. "Multiple Defender Effects: Synergistic Coral Defense by Mutualist Crustaceans." *Oecologia* 169 (4): 1095–1103. doi:10.1007/s00442-012-2275-2.

Ponciano, José Miguel, Mark L. Taper, Brian Dennis, and Subhash R. Lele. 2009. "Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning." *Ecology* 90 (2): 356–62. `http://www.jstor.org/stable/27650990`.

Rousset, François, and Jean-Baptiste Ferdy. 2014. "Testing Environmental and Genetic Effects in the Presence of Spatial Autocorrelation." *Ecography*, no–no. doi:10.1111/ecog.00566.

Rue, H., S. Martino, and N. Chopin. 2009. "Gaussian Models Using Integrated Nested Laplace Approximations (with Discussion)." *Journal of the Royal Statistical Society, Series B* 71 (2): 319–92.

Stroup, Walter W. 2014. "Rethinking the Analysis of Non-Normal Data in Plant and Soil Science." *Agronomy Journal* 106: 1–17. doi:10.2134/agronj2013.0342.

Sung, Yun Ju, and Charles J. Geyer. 2007. "Monte Carlo Likelihood Inference for Missing Data Models." *The Annals of Statistics* 35 (3): 990–1011. doi:10.1214/009053606000001389.