Review

Generalized linear mixed models: a practical guide for ecology and evolution

Benjamin M. Bolker[1], Mollie E. Brooks[1], Connie J. Clark[1], Shane W. Geange[2], John R. Poulsen[1], M. Henry H. Stevens[3] and Jada-Simone S. White[1]

[1] Department of Botany and Zoology, University of Florida, PO Box 118525, Gainesville, FL 32611-8525, USA
[2] School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
[3] Department of Botany, Miami University, Oxford, OH 45056, USA

How should ecologists and evolutionary biologists analyze nonnormal data that involve random effects? Nonnormal data such as counts or proportions often defy classical statistical procedures. Generalized linear mixed models (GLMMs) provide a more flexible approach for analyzing nonnormal data when random effects are present. The explosion of research on GLMMs in the last decade has generated considerable uncertainty for practitioners in ecology and evolution. Despite the availability of accurate techniques for estimating GLMM parameters in simple cases, complex GLMMs are challenging to fit and statistical inference such as hypothesis testing remains difficult. We review the use (and misuse) of GLMMs in ecology and evolution, discuss estimation and inference and summarize 'best-practice' data analysis procedures for scientists facing this challenge.

Generalized linear mixed models: powerful but challenging tools

Data sets in ecology and evolution (EE) often fall outside the scope of the methods taught in introductory statistics classes. Where basic statistics rely on normally distributed data, EE data are often binary (e.g. presence or absence of a species in a site [1], breeding success [2], infection status of individuals or expression of a genetic disorder [3]), proportions (e.g. sex ratios [4], infection rates [5] or mortality rates within groups) or counts (number of emerging seedlings [6], number of ticks on red grouse chicks [7] or clutch sizes of storks [2]). Where basic statistical methods try to quantify the exact effects of each predictor variable, EE problems often involve random effects, whose purpose is instead to quantify the variation among units. The most familiar types of random effect are the blocks in experiments or observational studies that are replicated across sites or times. Random effects also encompass variation among individuals (when multiple responses are measured per individual, such as survival of multiple offspring or sex ratios of multiple broods), genotypes, species and regions or time periods. Whereas geneticists and evolutionary biologists have long been interested in quantifying the magnitude of variation among genotypes [8–10], ecologists have more recently begun to appreciate the importance of random variation in space and time [11] or among individuals [12]. Theoretical studies emphasize the effects of variability on population dynamics [13,14]. In addition, estimating variability allows biologists to extrapolate statistical results to individuals or populations beyond the study sample.

Researchers faced with nonnormal data often try shortcuts such as transforming data to achieve normality and homogeneity of variance, using nonparametric tests or relying on the robustness of classical ANOVA to nonnormality for balanced designs [15]. They might ignore random effects altogether (thus committing pseudoreplication) or treat them as fixed factors [16]. However, such shortcuts can fail (e.g. count data with many zero values cannot be made normal by transformation). Even when they succeed, they might violate statistical assumptions (even nonparametric tests make assumptions, e.g. of homogeneity of variance across groups) or limit the scope of inference (one cannot extrapolate estimates of fixed effects to new groups).

Instead of shoehorning their data into classical statistical frameworks, researchers should use statistical approaches that match their data. Generalized linear mixed models (GLMMs) combine the properties of two statistical frameworks that are widely used in EE, linear mixed models (which incorporate random effects) and generalized linear models (which handle nonnormal data by using link functions and exponential family [e.g. normal, Poisson or binomial] distributions). GLMMs are the best tool for analyzing nonnormal data that involve random effects: all one has to do, in principle, is specify a distribution, link function and structure of the random effects. For example, in Box 1, we use a GLMM to quantify the magnitude of the genotype–environment interaction in the response of *Arabidopsis* to herbivory. To do so, we select a Poisson distribution with a logarithmic link (typical for count data) and specify that the total number of fruits per plant and the responses to fertilization and clipping could vary randomly across populations and across genotypes within a population.

However, GLMMs are surprisingly challenging to use even for statisticians. Although several software packages can handle GLMMs (Table 1), few ecologists and evolutionary biologists are aware of the range of options or of the possible pitfalls. In reviewing papers in EE since 2005

Corresponding author: Bolker, B.M. (bolker@ufl.edu).

0169-5347/$ – see front matter doi:10.1016/j.tree.2008.10.008

Glossary

Bayesian statistics: a statistical framework based on combining data with subjective prior information about parameter values in order to derive posterior probabilities of different models or parameter values.
Bias: inaccuracy of estimation, specifically the expected difference between an estimate and the true value.
Block random effects: effects that apply equally to all individuals within a group (experimental block, species, etc.), leading to a single level of correlation within groups.
Continuous random effects: effects that lead to between-group correlations that vary with distance in space, time or phylogenetic history.
Crossed random effects: multiple random effects that apply independently to an individual, such as temporal and spatial blocks in the same design, where temporal variability acts on all spatial blocks equally.
Exponential family: a family of statistical distributions including the normal, binomial, Poisson, exponential and gamma distributions.
Fixed effects: factors whose levels are experimentally determined or whose interest lies in the specific effects of each level, such as effects of covariates, differences among treatments and interactions.
Frequentist (sampling-based) statistics: a statistical framework based on computing the expected distributions of test statistics in repeated samples of the same system. Conclusions are based on the probabilities of observing extreme events.
Generalized linear models (GLMs): statistical models that assume errors from the exponential family; predicted values are determined by discrete and continuous predictor variables and by the link function (e.g. logistic regression, Poisson regression) (not to be confused with PROC GLM in SAS, which estimates general linear models such as classical ANOVA.).
Individual random effects: effects that apply at the level of each individual (i.e. 'blocks' of size 1).
Information criteria and information-theoretic statistics: a statistical framework based on computing the expected relative distance of competing models from a hypothetical true model.
Linear mixed models (LMMs): statistical models that assume normally distributed errors and also include both fixed and random effects, such as ANOVA incorporating a random effect.
Link function: a continuous function that defines the response of variables to predictors in a generalized linear model, such as logit and probit links. Applying the link function makes the expected value of the response linear and the expected variances homogeneous.
Markov chain Monte Carlo (MCMC): a Bayesian statistical technique that samples parameters according to a stochastic algorithm that converges on the posterior probability distribution of the parameters, combining information from the likelihood and the posterior distributions.
Maximum likelihood (ML): a statistical framework that finds the parameters of a model that maximizes the probability of the observed data (the likelihood). (See Restricted maximum likelihood.)
Model selection: any approach to determining the best of a set of candidate statistical models. Information-theoretic tools such as AIC, which also allow model averaging, are generally preferred to older methods such as stepwise regression.
Nested models: models that are subsets of a more complex model, derived by setting one or more parameters of the more complex model to a particular value (often zero).
Nested random effects: multiple random effects that are hierarchically structured, such as species within genus or subsites within sites within regions.
Overdispersion: the occurrence of more variance in the data than predicted by a statistical model.
Pearson residuals: residuals from a model which can be used to detect outliers and nonhomogeneity of variance.
Random effects: factors whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effects of each level. The parameters of random effects are the standard deviations of variation at a particular level (e.g. among experimental blocks). The precise definitions of 'fixed' and 'random' are controversial; the status of particular variables depends on experimental design and context [16,53].
Restricted maximum likelihood (REML): an alternative to ML that estimates the random-effect parameters (i.e. standard deviations) averaged over the values of the fixed-effect parameters; REML estimates of standard deviations are generally less biased than corresponding ML estimates.

found by Google Scholar, 311 out of 537 GLMM analyses (58%) used these tools inappropriately in some way (see online supplementary material). Here we give a broad but practical overview of GLMM procedures.

Whereas GLMMs themselves are uncontroversial, describing how to use them to analyze data necessarily touches on controversial statistical issues such as the debate over null hypothesis testing [17], the validity of stepwise regression [18] and the use of Bayesian statistics [19]. Others have thoroughly discussed these topics (e.g. [17–19]); we acknowledge the difficulty while remaining agnostic. We first discuss the estimation algorithms available for fitting GLMMs to data to find parameter estimates. We then describe the inferential procedures for constructing confidence intervals on parameters, comparing and selecting models and testing hypotheses with GLMMs. Finally, we summarize reasonable 'best practices' for using these techniques to answer ecological and evolutionary questions.

Estimation

Estimating the parameters of a statistical model is a key step in most statistical analyses. For GLMMs, these parameters are the fixed-effect parameters (effects of covariates, differences among treatments and interactions: in Box 1, these are the overall fruit set per individual and the effects of fertilization, clipping and their interaction on fruit set) and random-effect parameters (the standard deviations of the random effects: in Box 1, variation in fruit set, fertilization, clipping and interaction effects across genotypes and populations). Many modern statistical tools, including GLMM estimation, fit these parameters by maximum likelihood (ML). For simple analyses where the response variables are normal, all treatments have equal sample sizes (i.e. the design is balanced) and all random effects are nested effects, classical ANOVA methods based on computing differences of sums of squares give the same answers as ML approaches. However, this equivalence breaks down for more complex LMMs or for GLMMs: to find ML estimates, one must integrate likelihoods over all possible values of the random effects ([20,21] Box 2). For GLMMs this calculation is at best slow, and at worst (e.g. for large numbers of random effects) computationally infeasible.

Statisticians have proposed various ways to approximate the likelihood to estimate GLMM parameters, including pseudo- and penalized quasilikelihood (PQL [22–24]), Laplace approximations [25] and Gauss-Hermite quadrature (GHQ [26]), as well as Markov chain Monte Carlo (MCMC) algorithms [27] (Table 1). In all of these approaches, one must distinguish between standard ML estimation, which estimates the standard deviations of the random effects assuming that the fixed-effect estimates are precisely correct, and restricted maximum likelihood (REML) estimation, a variant that averages over some of the uncertainty in the fixed-effect parameters [28,29].

Box 1. A GLMM example: genotype-by-environment interaction in the response of *Arabidopsis* to herbivory

We used GLMMs to estimate gene-by-environment interaction in *Arabidopsis* response to simulated herbivory [54,55]. The fixed effects quantify the overall effects (across all genotypes) of fertilization and clipping; the random effects quantify the variation across genotypes and populations of the fixed-effect parameters. The random effects are a primary focus, rather than a nuisance variable.

Because the response variable (total fruits per individual) was count data, we started with a Poisson model (log link). The mean number of fruits per plant within genotype × treatment groups was sometimes <5, so we used Laplace approximation. Our 'full' model used fixed effects (nutrient + clipping + nutrient × clipping) and two sets of random effects that crossed these fixed effects with populations and genotypes within populations. Although populations were located within three larger regions, we ignored regional structure owing to insufficient replication. We also included two experimental design variables in all models, using fixed effects because of their small number of levels (both <4; Box 4). Laplace estimation methods for the full model converged easily.

The residuals indicated overdispersion, so we refitted the data with a quasi-Poisson model. Despite the large estimated scale parameter (10.8), exploratory graphs found no evidence of outliers at the level of individuals, genotypes or populations. We used quasi-AIC (QAIC), using one degree of freedom for random effects [49], for random-effect and then for fixed-effect model selection.

QAIC scores indicated that the model with all genotype-level random effects (nutrient, clipping and their interaction) and no population-level grouping was best; a model with population-level variation in overall fruit set was nearly as good (ΔQAIC = 0.6), and models with population-level variation in fertilization or clipping effects (but not both) were reasonable (ΔQAIC < 10). Because these models gave nearly identical fixed-effect estimates, model averaging was unnecessary. QAIC comparisons supported a strong average nutrient effect across all genotypes (threefold difference in fruit set), with weaker effects of clipping (50% decrease in fruit set, ΔQAIC = 1.9) and nutrient × clipping interaction (twofold increase, or compensating effects: ΔQAIC = 3.4).

The pattern of random effects (Figure I) indicated considerable heterogeneity across genotypes, with standard deviation ≈ 1 (at least as large as the fixed effects). Although the overall tendency for nutrients to allow plants to compensate for damage (fixed nutrient × clipping interaction) is weak, we infer strong gene-by-environment interaction at the level of individual genotypes.

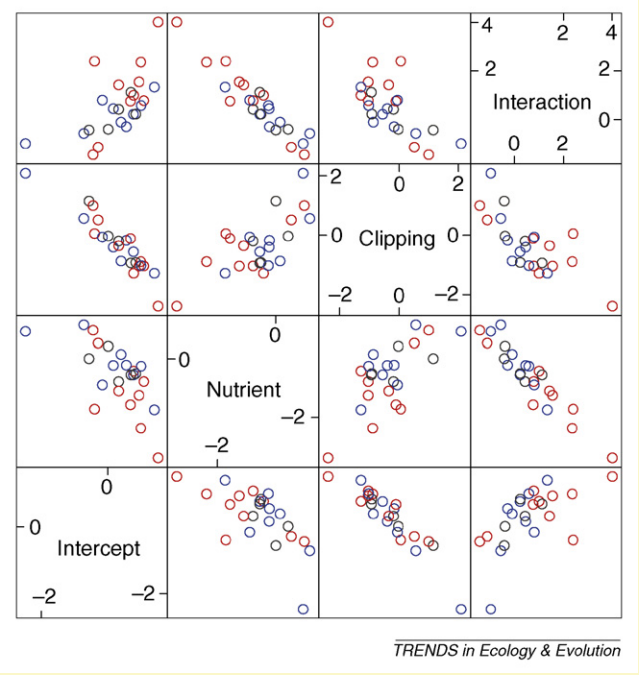

Figure I. Random effects of genotypes for each model parameter—differences of genotype-specific parameter values from the overall average. Diagonal panels give labels (intercept = log fruit set of control; nutrient = increase in log fruit set due to nutrient; clipping = decrease due to clipping; interaction = nutrient × clipping interaction) and scales for subplots. Color indicates region of origin.

ML underestimates random-effect standard deviations, except in very large data sets, but is more useful for comparing models with different fixed effects.

PQL is the simplest and most widely used GLMM approximation. Its implementation in widely available statistical packages has encouraged the use of GLMMs in many areas of EE, including behavioral and community ecology, conservation biology and quantitative and evolutionary genetics [30]. Unfortunately, PQL yields biased parameter estimates if the standard deviations of the

Table 1. Capabilities of different software packages for GLMM analysis: estimation methods, scope of statistical models that can be fitted and available inference methods

		Penalized quasilikelihood	Laplace	Gauss-Hermite quadrature	Crossed random effects	Wald χ^2 or Wald F tests	Degrees of freedom	MCMC sampling	Continuous spatial/ temporal correlation	Overdispersion
SAS	PROC GLIMMIX	✓	✓[a]	✓[a]	✓	✓	BW, S, KR		✓	QL
	PROC NLMIXED			✓		✓	BW, S, KR			Dist
R	glmmPQL	✓				✓	BW		✓	QL
	glmmML		✓	✓						
	glmer		✓	(✓)	✓			(✓)		QL
	glmmADMB		✓							Dist
	GLMM	✓			✓?	✓			✓	QL
GenStat/ ASREML			✓	✓	✓			✓		Dist
AD Model Builder		✓	✓		✓					✓
HLM				✓						
GLLAAMM (Stata)								✓		Dist
WinBUGS					✓			✓		

Abbreviations: BW, between–within; dist, specified distribution (e.g. negative binomial); KR, Kenward-Roger; QL, quasilikelihood; S, Satterthwaite.
[a]Version 9.2 only.

Box 2. Estimation details: evaluating GLMM likelihoods

Consider data x with a single random effect θ (e.g. the difference of blocks from the overall mean) with variance σ^2 (e.g. the variance among blocks) and fixed-effect parameter μ (e.g. the expected difference between two treatments). The overall likelihood is $\int P(\theta|\sigma^2)^2) L(x|\mu,\theta) d\theta$: the first term $[P(\theta|\sigma^2)]$ gives the probability of drawing a particular block value θ from the (normally distributed) block distribution, while the second term $[L(x|\mu,\theta)]$ gives the probability of observing the data given the treatment effect and the particular block value. Integrating computes the average likelihood across all possible block values, weighted by their probability [28]. Procedures for GLMM parameter estimation approximate the likelihood in several different ways (Table I):

- Penalized quasilikelihood alternates between (i) estimating fixed parameters by fitting a GLM with a variance–covariance matrix based on an LMM fit and (ii) estimating the variances and covariances by fitting an LMM with unequal variances calculated from the previous GLM fit. Pseudolikelihood, a closely related technique, estimates the variances in step ii differently and estimates a scale parameter to account for overdispersion (some authors use these terms interchangeably).
- The Laplace method approximates the likelihood by assuming that the distribution of the likelihood (*not* the distribution of the data) is approximately normal, making the likelihood function quadratic on the log scale and allowing the use of a second-order Taylor expansion.
- Gauss-Hermite quadrature approximates the likelihood by picking optimal subdivisions at which to evaluate the integrand. Adaptive GHQ incorporates information from an initial fit to increase precision.
- Markov chain Monte Carlo algorithms sample sequentially from random values of the fixed-effect parameters, the levels of the random effects (θ in the example above) and random-effect parameters (σ^2 above), in a way that converges on the distribution of these values.

These procedures are unnecessary for linear mixed models, although mistaken use of GLMM techniques to analyze LMMs is widespread in the literature (see online supplement).

Table I. Techniques for GLMM parameter estimation, their advantages and disadvantages and the software packages that implement them

Technique	Advantages	Disadvantages	Software
Penalized quasilikelihood	Flexible, widely implemented	Likelihood inference inappropriate; biased for large variance or small means	PROC GLIMMIX (SAS), GLMM (Genstat), glmmPQL (R), glmer (R)
Laplace approximation	More accurate than PQL	Slower and less flexible than PQL	PROC GLIMMIX [56], glmer (R), glmm.admb (R), AD Model Builder, HLM
Gauss-Hermite quadrature	More accurate than Laplace	Slower than Laplace; limited to 2–3 random effects	PROC GLIMMIX [56], PROC NLMIXED (SAS), glmer (R), glmmML (R)
Markov chain Monte Carlo	Highly flexible, arbitrary number of random effects; accurate	Very slow, technically challenging, Bayesian framework	WinBUGS, JAGS, MCMCpack, (R), AD Model Builder

random effects are large, especially with binary data (i.e. binomial data with a single individual per observation) [31,32]. Statisticians have implemented several improved versions of PQL, but these are not available in the most common software packages ([32,33]). As a rule of thumb, PQL works poorly for Poisson data when the mean number of counts per treatment combination is less than five, or for binomial data where the expected numbers of successes and failures for each observation are both less than five (which includes binary data) [30]. Nevertheless, our literature review found that 95% of analyses of binary responses (n = 205), 92% of Poisson responses with means less than 5 (n = 48) and 89% of binomial responses with fewer than 5 successes per group (n = 38) used PQL.

Another disadvantage of PQL is that it computes a quasilikelihood rather than a true likelihood. Many statisticians feel that likelihood-based methods should not be used for inference (e.g. hypothesis testing, AIC ranking) with quasilikelihoods (see Inference section below [26]).

Two more accurate approximations are available [25,30]. As well as reducing bias, Laplace approximation (Box 2 [25]) approximates the true GLMM likelihood rather than a quasilikelihood, allowing the use of likelihood-based inference. Gauss-Hermite quadrature [26] is more accurate still, but is slower than Laplace approximation. Because the speed of GHQ decreases rapidly with increasing numbers of random effects, it is not feasible for analyses with more than two or three random factors.

In contrast to methods that explicitly integrate over random effects to compute the likelihood, MCMC methods generate random samples from the distributions of parameter values for fixed and random effects. MCMC is usually used in a Bayesian framework, which incorporates prior information based on previous knowledge about the parameters or specifies uninformative (weak) prior distributions to indicate lack of knowledge. Inference is based on summary statistics (mean, mode, quantiles, etc.) of the posterior distribution, which combines the prior distribution with the likelihood [34]. Bayesian MCMC gives similar answers to maximum-likelihood approaches when data sets are highly informative and little prior knowledge is assumed (i.e. when the priors are weak). Unlike the methods discussed above, MCMC methods extend easily to consider multiple random effects [27], although large data sets are required. In addition to its Bayesian flavor (which might deter some potential users), MCMC involves several potentially difficult technical details, including making sure that the statistical model is well posed; choosing appropriate priors [35]; choosing efficient algorithms for large problems [36]; and assessing when chains have run long enough for reliable estimation [37–39]. Statisticians are also developing alternative tools that exploit the computational advantages of MCMC within a frequentist framework [40,41], but these approaches have not been widely tested.

Although many estimation tools are only available in a few statistics packages, or are difficult to use, the situation

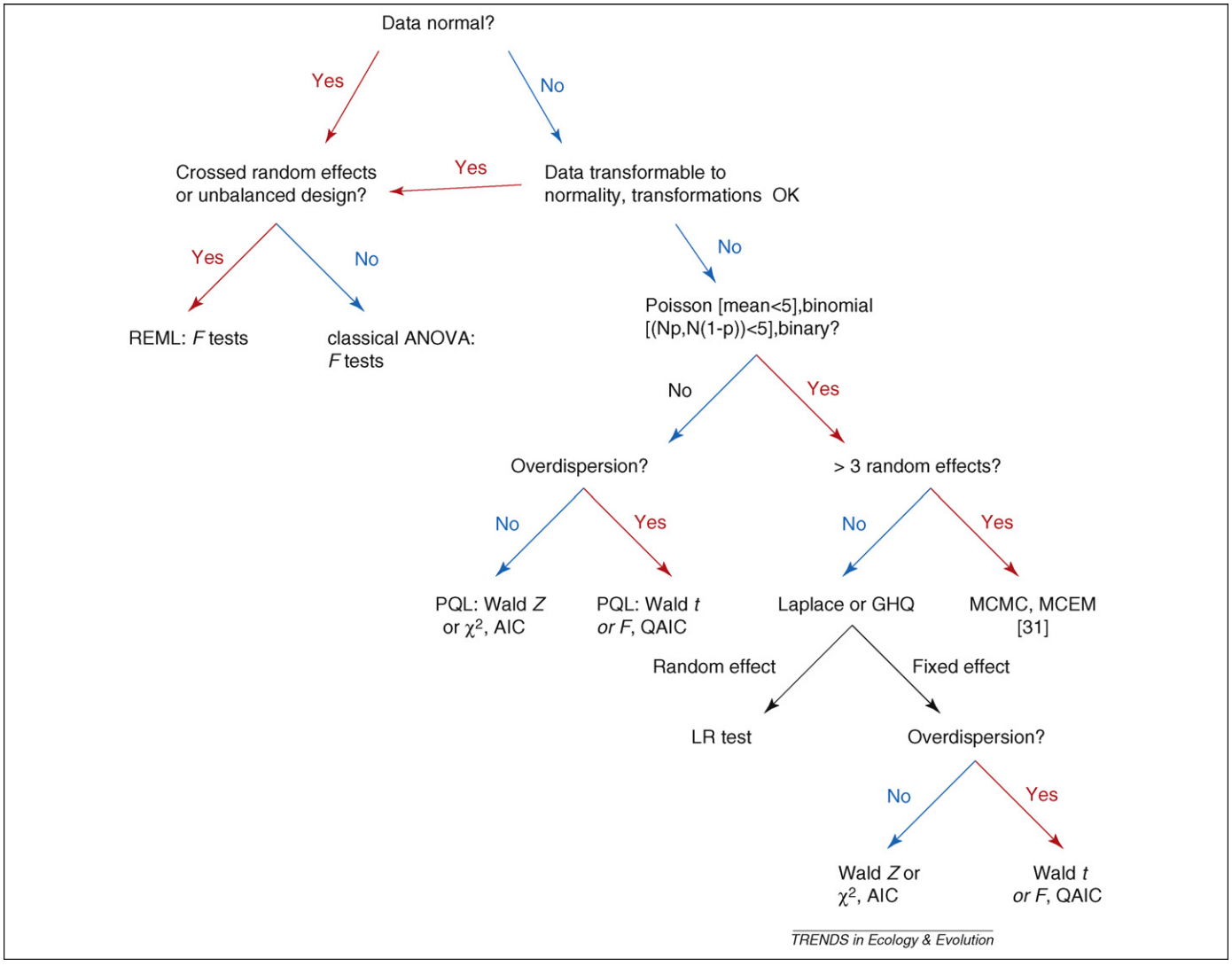

Figure 1. Decision tree for GLMM fitting and inference. Conditions on the Poisson and binomial distributions along the right branch refer to penalized quasilikelihood (PQL) rules of thumb [30]: to use PQL, Poisson distributions should have mean > 5 and binomial distributions should have the minimum of the number of successes and failures > 5. MCEM = Monte Carlo expectation-maximization [40].

is gradually improving as software developers and publishers improve their offerings. Which estimation technique is most useful in a given situation depends on the complexity of the model, as well as computation time, availability of software and applicability of different inference methods (Figure 1).

Inference

After estimating parameter values for GLMMs, the next step is statistical inference: that is, drawing statistical and biological conclusions from the data by examining the estimates and their confidence intervals, testing hypotheses, selecting the best model(s) and evaluating differences in goodness of fit among models. We discuss three general types of inference: hypothesis testing, model comparison and Bayesian approaches. Frequentist hypothesis testing compares test statistics (e.g. F statistics in ANOVA) to their expected distributions under the null hypothesis, estimating a p value to determine whether one can reject the null hypothesis. Model selection, by contrast, compares fits of candidate models. One can select models either by using hypothesis tests (i.e. testing simpler nested models against more complex models) [42] or by using information-theoretic approaches, which use measures of expected predictive power to rank models or average their predictions [43]. Bayesian methods have the same general scope as frequentist or information-theoretic approaches, but differ in their philosophical underpinnings as well as in the specific procedures used.

Hypothesis testing

Wald Z, χ^2, t and F tests for GLMMs test a null hypothesis of no effect by scaling parameter estimates or combinations of parameters by their estimated standard errors and comparing the resulting test statistic to zero [44]. Wald Z and χ^2 tests are only appropriate for GLMMs without overdispersion, whereas Wald t and F tests account for the uncertainty in the estimates of overdispersion [29]. This uncertainty depends on the number of residual degrees of freedom, which can be very difficult to calculate because the effective number of parameters used by a random effect lies somewhere between 1 (i.e. a single standard deviation

Box 3. Inference details

Drawing inferences (e.g. testing hypotheses) from the results of GLMM analyses can be challenging, and in some cases statisticians still disagree on appropriate procedures (Table I). Here we highlight two particular challenges, boundary effects and calculating degrees of freedom.

Boundary effects

Many tests assume that the null values of the parameters are not on the boundary of their allowable ranges. In particular, the null hypothesis for random effects ($\sigma = 0$) violates this assumption, because standard deviations must be ≥ 0 [45]. Likelihood ratio tests that compare the change in deviance between nested models that differ by v random-effect terms against a χ^2 distribution with v degrees of freedom (χ^2_v) are conservative, increasing the risk of type II errors. Mixtures of χ^2_v and χ^2_{v-1} distributions are appropriate in simple cases [57–59]; for a single variance parameter ($v = 1$), this is equivalent to dividing the standard χ^2_1 p value by 2 [29]. Information-theoretic approaches suffer from analogous problems [48,60].

Calculating degrees of freedom

The degrees of freedom (df) for random effects, needed for Wald t or F tests or AIC_c, must be between 1 and $N - 1$ (where N is the number of random-effect levels). Software packages vary enormously in their approach to computing df [61]. The simplest approach (the default in SAS) uses the minimum number of df contributed by random effects that affect the term being tested [29]. The Satterthwaite and Kenward-Roger (KR) approximations [29,62] use more complicated rules to approximate the degrees of freedom and adjust the standard errors. KR, only available in SAS, generally performs best (at least for LMMs [63]). In our literature review, most SAS analyses (63%, n = 102) used the default method (which is 'at best approximate, and can be unpredictable' [64]). An alternative approach uses the hat matrix, which can be derived from GLMM estimates. The sample size n minus the trace t (i.e. the sum of the diagonal elements) of the hat matrix provides an estimate of the residual degrees of freedom [43,51]. If the adjusted residual df are >25, these details are less important.

Accounting for boundary effects and computing appropriate degrees of freedom is still difficult. Researchers should use appropriate corrections when they are available, and understand the biases that occur in cases where such corrections are not feasible (e.g. ignoring boundary effects makes tests of random effects conservative).

Table I. Techniques for GLMM inferences, their advantages and disadvantages and the software packages that implement them

Method	Advantages	Disadvantages	Software
Wald tests (Z, χ^2, t, F)	Widely available, flexible, OK for quasilikelihood (QL)	Boundary issues; poor for random effects; t and F require residual df	GLIMMIX, NLMIXED (SAS), glmmPQL (R)
Likelihood ratio test	Better than Wald tests for random effects	Bad for fixed effects without large sample sizes; boundary effects; inappropriate for QL	NLMIXED (SAS), lme4 (R)
Information criteria	Avoids stepwise procedures; provides model weights and averaging; QAIC applies to overdispersed data	Boundary effects; no p value; requires residual df estimate for AIC_c	GLIMMIX, NLMIXED (SAS), lme4 (R)
Deviance information criterion	Automatically penalizes model complexity	Requires MCMC sampling	WinBUGS

parameter) and $N - 1$ (i.e. one parameter for each additional level of the random effect; [29] Box 3). For random effects, these tests (in common with several other GLMM inference tools) suffer from boundary effects because the null values of the parameters lie at the edge of their feasible range ([45] Box 3): that is, the standard deviations can only be greater and not less than their null-hypothesis value of zero.

The likelihood ratio (LR) test determines the contribution of a single (random or fixed) factor by comparing the fit (measured as the deviance, i.e. −2 times the log-likelihood ratio) for models with and without the factor, namely nested models. Although widely used throughout statistics, the LR test is not recommended for testing fixed effects in GLMMs, because it is unreliable for small to moderate sample sizes (note that LR tests on fixed effects, or any comparison of models with different fixed effects, also require ML, rather than REML, estimates) [28]. The LR test is only adequate for testing fixed effects when both the ratio of the total sample size to the number of fixed-effect levels being tested [28] and the number of random-effect levels (blocks) [44,46] are large. We have found little guidance and no concrete rules of thumb in the literature on this issue, and would recommend against using the LR test for fixed effects unless the total sample size and numbers of blocks are very large. The LR test is generally appropriate for inference on random factors, although corrections are needed to address boundary problems similar to those of the Wald tests [28,45]. In general, because Wald tests make stronger assumptions, LR tests are preferred for inference on random effects [47] (Figure 1).

Model selection and averaging

LR tests can assess the significance of particular factors or, equivalently, choose the better of a pair of nested models, but some researchers have criticized model selection via such pairwise comparisons as an abuse of hypothesis testing [18,43]. Information-theoretic model selection procedures, by contrast, allow comparison of multiple, nonnested models. The Akaike information criterion (AIC) and related information criteria (IC) use deviance as a measure of fit, adding a term to penalize more complex models (i.e. greater numbers of parameters). Rather than estimating p values, information-theoretic methods estimate statistics that quantify the magnitude of difference between models in expected predictive power, which one can then assess using rules of thumb [43]. ICs also provide a natural basis for averaging parameter estimates and predictions across models, which can provide better estimates as well as confidence intervals that correctly account for model uncertainty [17]. Variants of AIC are useful when sample sizes are small (AIC_c), when the data are overdispersed (quasi-AIC, QAIC) or when one wants to identify the number of parameters in a 'true' model (Bayesian or Schwarz information criterion, BIC) [43]. The main concerns with using AIC for GLMMs (boundary effects [48] and estimation of degrees of freedom for random effects [49]) mirror those for classical statistical tests (Box 3).

Box 4. Procedures: creating a full model

Here we outline a general framework for constructing a full (most complex) model, the first step in GLMM analysis. Following this process, one can then evaluate parameters and compare submodels as described in the main text and in Figure 1.

1. Specify fixed (treatments or covariates) and random effects (experimental, spatial or temporal blocks, individuals, etc.). Include only important interactions. Restrict the model *a priori* to a feasible level of complexity, based on rules of thumb (>5–6 random-effect levels per random effect and >10–20 samples per treatment level or experimental unit) and knowledge of adequate sample sizes gained from previous studies [64,65].
2. Choose an error distribution and link function (e.g. Poisson distribution and log link for count data, binomial distribution and logit link for proportion data).
3. Graphical checking: are variances of data (transformed by the link function) homogeneous across categories? Are responses of transformed data linear with respect to continuous predictors? Are there outlier individuals or groups? Do distributions within groups match the assumed distribution?
4. Fit fixed-effect GLMs both to the full (pooled) data set and within each level of the random factors [28,50]. Estimated parameters should be approximately normally distributed across groups (group-level parameters can have large uncertainties, especially for groups with small sample sizes). Adjust model as necessary (e.g. change link function or add covariates).
5. Fit the full GLMM.
   Insufficient computer memory or *too slow:* reduce model complexity. If estimation succeeds on a subset of the data, try a more efficient estimation algorithm (e.g. PQL if appropriate).
   Failure to converge (warnings or errors): reduce model complexity or change optimization settings (make sure the resulting answers make sense). Try other estimation algorithms.
   Zero variance components or singularity (warnings or errors): check that the model is properly defined and identifiable (i.e. all components can theoretically be estimated). Reduce model complexity.
   Adding information to the model (additional covariates, or new groupings for random effects) can alleviate problems, as will centering continuous covariates by subtracting their mean [50]. If necessary, eliminate random effects from the full model, dropping (i) terms of less intrinsic biological interest, (ii) terms with very small estimated variances and/or large uncertainty, or (iii) interaction terms. (Convergence errors or zero variances could indicate insufficient data.)
6. Recheck assumptions for the final model (as in step 3) and check that parameter estimates and confidence intervals are reasonable (gigantic confidence intervals could indicate fitting problems). The magnitude of the standardized residuals should be independent of the fitted values. Assess overdispersion (the sum of the squared Pearson residuals should be χ^2 distributed [66,67]). If necessary, change distributions or estimate a scale parameter. Check that a full model that includes dropped random effects with small standard deviations gives similar results to the final model. If different models lead to substantially different parameter estimates, consider model averaging.

Bayesian approaches

Bayesian approaches to GLMM inference offer several advantages over frequentist and information-theoretic methods [50]. First, MCMC provides confidence intervals on GLMM parameters (and hence tests of whether those parameters could plausibly equal zero) in a way that naturally averages over the uncertainty in both the fixed- and random-effect parameters, avoiding many of the difficult approximations used in frequentist hypothesis testing. Second, Bayesian techniques define posterior model probabilities that automatically penalize more complex models, providing a way to select or average over models. Because these probabilities can be very difficult to compute, Bayesian analyses typically use two common approximations, the Bayesian (BIC) and deviance (DIC) information criteria [51]. The BIC is similar to the AIC, and similarly requires an estimate of the number of parameters (Box 3). The DIC makes weaker assumptions, automatically estimates a penalty for model complexity and is automatically calculated by the WinBUGS program (http://www.mrc-bsu.cam.ac.uk/bugs). Despite uncertainty among statisticians about its properties [51], the DIC is rapidly gaining popularity in ecological and evolutionary circles.

One can also use Bayesian approaches to compute confidence intervals for model parameters estimated by frequentist methods [52] by using a specialized MCMC algorithm that samples from the posterior distribution of the parameters (assuming uninformative priors). This approach represents a promising alternative that takes uncertainty in both fixed- and random-effect parameters into account, capitalizes on the computational efficiency of frequentist approaches and avoids the difficulties of estimating degrees of freedom for F tests, but it has only been implemented very recently [52].

Procedures

Given all of this information, how should one actually use GLMMs to analyze data (Box 4)? Unfortunately, we cannot recommend a single, universal procedure because different methods are appropriate for different problems (see Figure 1) and, as made clear by recent debates [42], how one analyzes data depends strongly on one's philosophical approach (e.g. hypothesis testing versus model selection, frequentist versus Bayesian). In any case, we strongly recommend that researchers proceed with caution by making sure that they have a good understanding of the basics of linear and generalized mixed models before taking the plunge into GLMMs, and by respecting the limitations of their data.

After constructing a full model (Box 4), one must choose among philosophies of inference. The first option is classical backward stepwise regression using the LR test to test random effects and Wald χ^2 tests, Wald F tests or MCMC sampling to test fixed effects, discarding effects that do not differ significantly from zero. Whereas statisticians strongly discourage automatic stepwise regression with many potential predictors, disciplined hypothesis testing for small amounts of model reduction is still considered appropriate in some situations [28].

Alternatively, information-theoretic tools can select models of appropriate complexity [43]. This approach finds the model with the highest estimated predictive power, without data snooping, assuming that we can accurately estimate the number of parameters (i.e. degrees of free-

dom) for random effects [49]. Ideally, rather than selecting the 'best' model, one would average across all reasonably well fitting models (e.g. $\Delta AIC < 10$), using either IC or Bayesian tools [43], although the additional complexity of this step could be unnecessary if model predictions are similar or if qualitative understanding rather than quantitative prediction is the goal of the study.

Finally, one could assume that all of the effects included in the full model are really present, whether statistically significant or not. One would then estimate parameters and confidence intervals from the full model, avoiding any data snooping problems but paying the penalty of larger variance in predictions; many Bayesian analyses, especially those of large data sets where loss of precision is less important, take this approach [50].

It is important to distinguish between random effects as a nuisance (as in classical blocked experimental designs) and as a variable of interest (as in many evolutionary genetic studies, or in ecological studies focused on heterogeneity). If random effects are part of the experimental design, and if the numerical estimation algorithms do not break down, then one can choose to retain all random effects when estimating and analyzing the fixed effects. If the random effects are a focus of the study, one must choose between retaining them all, selecting some by stepwise or all-model comparison, or averaging models.

Conclusion

Ecologists and evolutionary biologists have much to gain from GLMMs. GLMMs allow analysis of blocked designs in traditional ecological experiments with count or proportional responses. By incorporating random effects, GLMMs also allow biologists to generalize their conclusions to new times, places and species. GLMMs are invaluable when the random variation is the focus of attention, particularly in studies of ecological heterogeneity or the heritability of discrete characters.

In this review, we have encouraged biologists to choose appropriate tools for GLMM analyses, and to use them wisely. With the rapid advancement of statistical tools, many of the challenges emphasized here will disappear, leaving only the fundamental challenge of posing feasible biological questions and gathering enough data to answer them.

Acknowledgements

We would like to thank Denis Valle, Paulo Brando, Jim Hobert, Mike McCoy, Craig Osenberg, Will White, Ramon Littell and members of the R-sig-mixed-models mailing list (Douglas Bates, Ken Beath, Sonja Greven, Vito Muggeo, Fabian Scheipl and others) for useful comments. Josh Banta and Massimo Pigliucci provided data and guidance on the *Arabidopsis* example. S.W.G. was funded by a New Zealand Fulbright–Ministry of Research, Science and Technology Graduate Student Award.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tree.2008.10.008.

References

1 Milsom, T. *et al.* (2000) Habitat models of bird species distribution: an aid to the management of coastal grazing marshes. *J. Appl. Ecol.* 37, 706–727
2 Vergara, P. and Aguirre, J.I. (2007) Arrival date, age and breeding success in white stork *Ciconia ciconia*. *J. Avian Biol.* 38, 573–579
3 Pawitan, Y. *et al.* (2004) Estimation of genetic and environmental factors for binary traits using family data. *Stat. Med.* 23, 449–465
4 Kalmbach, E. *et al.* (2001) Increased reproductive effort results in male-biased offspring sex ratio: an experimental study in a species with reversed sexual size dimorphism. *Proc. Biol. Sci.* 268, 2175–2179
5 Smith, A. *et al.* (2006) A role for vector-independent transmission in rodent trypanosome infection? *Int. J. Parasitol.* 36, 1359–1366
6 Jinks, R.L. *et al.* (2006) Direct seeding of ash (*Fraxinus excelsior* L.) and sycamore (*Acer pseudoplatanus* L.): the effects of sowing date, pre-emergent herbicides, cultivation, and protection on seedling emergence and survival. *For. Ecol. Manage.* 237, 373–386
7 Elston, D.A. *et al.* (2001) Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* 122, 563–569
8 Gilmour, A.R. *et al.* (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72, 593–599
9 Kruuk, L.E.B. *et al.* (2002) Antler size in red deer: heritability and selection but no evolution. *Evolution* 56, 1683–1695
10 Wilson, A.J. *et al.* (2006) Environmental coupling of selection and heritability limits evolution. *PLoS Biol.* 4, e216
11 Chesson, P. (2000) Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* 31, 343–366
12 Melbourne, B.A. and Hastings, A. (2008) Extinction risk depends strongly on factors contributing to stochasticity. *Nature* 454, 100–103
13 Fox, G.A. and Kendall, B.E. (2002) Demographic stochasticity and the variance reduction effect. *Ecology* 83, 1928–1934
14 Pfister, C.A. and Stevens, F.R. (2003) Individual variation and environmental stochasticity: implications for matrix model predictions. *Ecology* 84, 496–510
15 Quinn, G.P. and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists.* Cambridge University Press
16 Crawley, M.J. (2002) *Statistical Computing: An Introduction to Data Analysis Using S-PLUS.* John Wiley & Sons
17 Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.* 19, 101–108
18 Whittingham, M.J. *et al.* (2006) Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.* 75, 1182–1189
19 Ellison, A.M. (2004) Bayesian inference in ecology. *Ecol. Lett.* 7, 509–520
20 Browne, W.J. and Draper, D. (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* 1, 473–514
21 Lele, S.R. (2006) Sampling variability and estimates of density dependence: a composite-likelihood approach. *Ecology* 87, 189–202
22 Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika* 78, 719–727
23 Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simulation* 48, 233–243
24 Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25
25 Raudenbush, S.W. *et al.* (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* 9, 141–157
26 Pinheiro, J.C. and Chao, E.C. (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J. Comput. Graph. Statist.* 15, 58–81
27 Gilks, W.R. *et al.* (1996) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (Gilks, W.R., ed.), pp. 1–19, Chapman and Hall
28 Pinheiro, J.C. and Bates, D.M. (2000) *Mixed-Effects Models in S and S-PLUS.* Springer
29 Littell, R.C. *et al.* (2006) *SAS for Mixed Models.* (2nd edn), SAS Publishing
30 Breslow, N.E. (2004) Whither PQL? In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data* (Lin, D.Y. and Heagerty, P.J., eds), pp. 1–22, Springer
31 Rodriguez, G. and Goldman, N. (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *J. R. Stat. Soc. Ser. A Stat. Soc.* 164, 339–355

32 Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Stat. Soc. Ser. A Stat. Soc.* 159, 505–513
33 Lee, Y. and Nelder, J.A. (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* 88, 987–1006
34 McCarthy, M. (2007) *Bayesian Methods for Ecology*. Cambridge University Press
35 Berger, J. (2006) The case for objective Bayesian analysis. *Bayesian Anal.* 1, 385–402
36 Carlin, B.P. *et al.* (2006) Elements of hierarchical Bayesian inference. In *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications* (Clark, J.S. and Gelfand, A.E., eds), pp. 3–24, Oxford University Press
37 Cowles, M.K. and Carlin, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* 91, 883–904
38 Brooks, S.P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* 7, 434–455
39 Paap, R. (2002) What are the advantages of MCMC based inference in latent variable models? *Stat. Neerl.* 56, 2–22
40 Booth, J.G. and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B Methodological* 61, 265–285
41 Lele, S.R. *et al.* (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* 10, 551–563
42 Stephens, P.A. *et al.* (2005) Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* 42, 4–12
43 Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag
44 Agresti, A. (2002) *Categorical Data Analysis*. Wiley-Interscience
45 Molenberghs, G. and Verbeke, G. (2007) Likelihood ratio, score, and Wald tests in a constrained parameter space. *Am. Stat.* 61, 22–27
46 Demidenko, E. (2004) *Mixed Models: Theory and Applications*. Wiley-Interscience
47 Scheipl, F. *et al.* (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Stat. Data Anal.* 52, 3283–3299
48 Greven, S. (2008) *Non-Standard Problems in Inference for Additive and Linear Mixed Models*. Cuvillier Verlag
49 Vaida, F. and Blanchard, S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika* 92, 351–370
50 Gelman, A. and Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press
51 Spiegelhalter, D.J. *et al.* (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* 64, 583–640
52 Baayen, R.H. *et al.* (2008) Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412
53 Gelman, A. (2005) Analysis of variance: why it is more important than ever. *Ann. Stat.* 33, 1–53
54 Banta, J.A. *et al.* (2007) Evidence of local adaptation to coarse-grained environmental variation in *Arabidopsis thaliana*. *Evolution* 61, 2419–2432
55 Banta, J.A. (2008) Tolerance to apical meristem damage in *Arabidopsis thaliana* (Brassicaceae): a closer look and the broader picture. PhD dissertation, Stony Brook University
56 Schabenberger, O. (2007) Growing up fast: SAS® 9.2 enhancements to the GLIMMIX procedure. SAS Global Forum 2007, 177 (www2.sas.com/proceedings/forum2007/177-2007.pdf)
57 Self, S.G. and Liang, K-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* 82, 605–610
58 Stram, D.O. and Lee, J.W. (1994) Variance components testing in the longitudinal fixed effects model. *Biometrics* 50, 1171–1177
59 Goldman, N. and Whelan, S. (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 17, 975–978
60 Dominicus, A. *et al.* (2006) Likelihood ratio tests in behavioral genetics: problems and solutions. *Behav. Genet.* 36, 331–340
61 Aukema, B.H. *et al.* (2005) Quantifying sources of variation in the frequency of fungi associated with spruce beetles: implications for hypothesis testing and sampling methodology in bark beetle-symbiont relationships. *For. Ecol. Manage.* 217, 187–202
62 Schaalje, G.B. *et al.* (2001) Approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. *SUGI (SAS User's Group International)* 26, 262 (www2.sas.com/proceedings/sugi26/p262-26.pdf)
63 Schaalje, G. *et al.* (2002) Adequacy of approximations to distributions of test statistics in complex mixed linear models. *J. Agric. Biol. Environ. Stat.* 7, 512–524
64 Gotelli, N.J. and Ellison, A.M. (2004) *A Primer of Ecological Statistics*. Sinauer Associates
65 Harrell, F.J. (2001) *Regression Modeling Strategies*. Springer
66 Lindsey, J.K. (1997) *Applying Generalized Linear Models*. Springer
67 Venables, W. and Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer