Notes on inference and model selection with mixed effect models

Marco D. Visser*[1,2]

[1]Departments of Experimental Plant Ecology and Animal Ecology & Ecophysiology, Radboud University Nijmegen, the Netherlands
[2]Smithsonian Tropical Research Institute, Panama

March 12, 2015

Contents

1 Introduction

Mixed models (MM) are a popular tool in ecology today, one that is rapidly gaining popularity as MM have become easy to implement in most software packages (SAS, SPSS, R:lme4). The popularity of MM is likely largely due to the fact that ecological datasets often violate the assumptions of classical

*m.visser@science.ru.nl

statistical tests. There is also increased interest in directly estimating variance (between individuals, or in space and time) as theoretical studies emphasize the effects of variability on e.g. population dynamics (Pfister & Stevens, 2003). However, recent studies have shown the majority of studies in ecology (58 - 95%) used these tools inappropriately (Bolker *et al.*, 2009). Much is still unknown about mixed models, this document aims to summarize some of the key issues when trying to apply model selection and inference to MM, without going into great detail. The document assumes some knowledge of MM.

2 How Maximum Likelihood estimate are approximated. A key consideration.

To obtain Maximum Likelihood estimated for mixed models with random effects one must integrate likelihoods over all possible values of the random effects. For instance, if we are studying a system of organisms (e.g. seedlings, owlets, daphnia) and we were interested in the variation in survival over time, as well as the "classical" mean survival. The system could be described by:

$$S_t \sim bin(p, N_{t-1}) \tag{2.1}$$

Where S_t are the amount of surviving individuals at time t, from an original population of N_{t-1} and p is a random variable distributed as $p \sim beta(\alpha, \beta)$. The likelihood of observing a set of S survivors, from N individuals over T years, given the parameters *alpha* and *beta* would be:

$$L(\alpha, \beta \mid S, N) = \prod_{t=1}^{T} \left[\int_0^1 beta(p \mid \alpha, \beta) bin(S_t, N_{t-1} \mid p) dp \right] \tag{2.2}$$

This example of integrating over all values of the the "random effect" p to obtain the MLE for *alpha* and *beta* is one of the few cases where an analytical solution exists (called the beta-binomial). However, in most cases, no analytical solution exists and integration must be done numerically. Even for simple problems this quickly becomes infeasible. For these reasons statisticians have come up ways to approximate the MLE of model parameters including random effects. These techniques include:

1. Pseudo and penalized quasi-likelihoods [PQL]
2. Laplace Approximations [LA]
3. Guasse-Hermite quadrature [GHQ]
4. Monte Carlo Markov Chain methods [MCMC]

All of the above can again be distinguished between standard ML estimation, in which the fixed effect parameters are assumed to be precisely correct when

estimating the random effects (as above in 2.2), or restricted maximum likelihood (REML) which averages over uncertainty in the fixed effects (Pinheiro & Bates, 2000). It is good to consider the precise method used in approximating the MLE, as this has serious consequences for model inference and selection.

3 Mixed effect models: benefits

Mixed effect models are essentially multi-level or hierarchical models. Many researchers view multi-level models as superior stating that they are almost always an improvement. Here is a list of commonly associated benefits (Gelman, 2006):

1. Models in which you pool all data (e.g. average over groups), are likely to underfit the data, while model where you don't pool the data (e.g. fit one model per group) tend to overfit the data.

2. Multilevel models outperform classical regression in predictive accuracy. Cross validation studies show that multilevel models, by allowing shrinkage, have lower mean squared error for prediction.

3. In inference, correlations can lead to erroneous conclusions with classical approaches. This happens when we confound variables at different levels.

The following simulation study illustrates some of the benefits of the hierarchical nature of mixed effect models. Lets assume we sampled data (Y) from a population with 20 distinct groups. Within each group, Y related to X, as follows:

$$\begin{aligned} X &\sim Normal(\mu = X_g, \sigma = 0.01) \\ Y &\sim Normal(0, 1) + Normal(0.5X, 0.01) \end{aligned} \tag{3.1}$$

Where X_g is the group average. While the group mean values are uniformly distributed.

$$X_g \sim Uniform(0, 4) \tag{3.2}$$

We also have a common unbalanced design where some groups have more samples than others, here 5 groups have 2000 samples and the others 100. Our simulation has a clear multi-level structure of a within group trends, which does not exist on the between group means. The graph below shows the performance of the classic approaches and a mixed effect model.

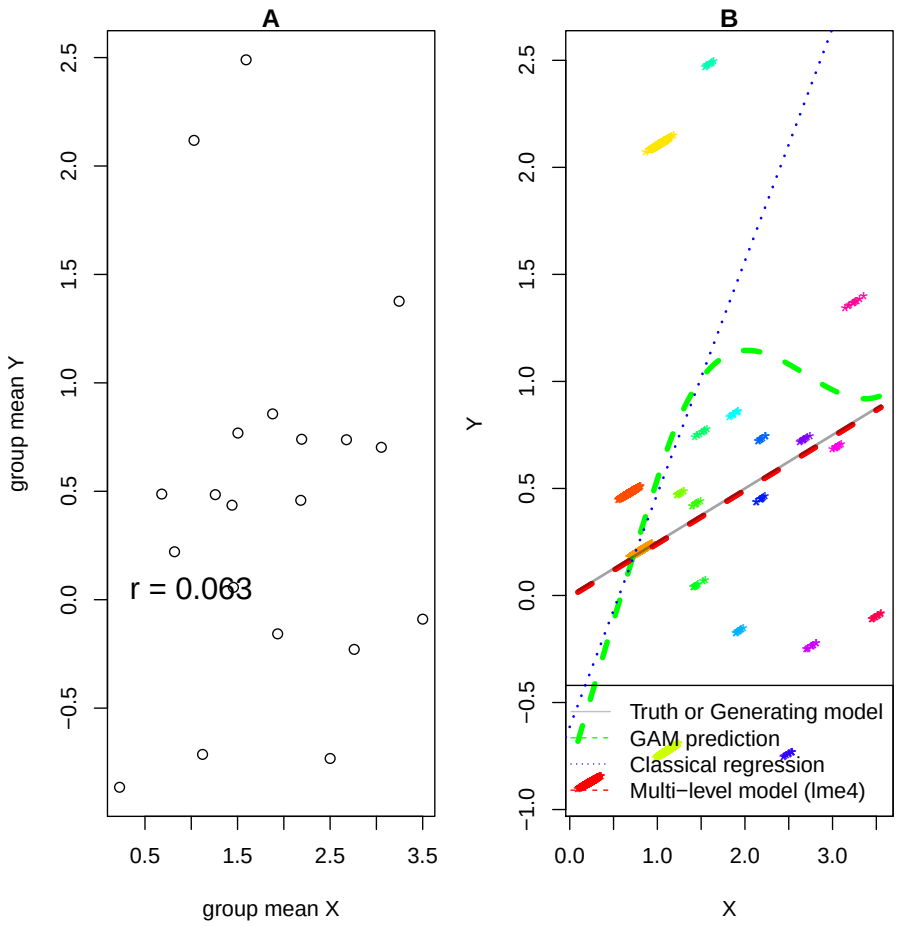

Figure 3.1: In Panel A, we correlate the groups means and find no relationship (r=0.06,p=0.53). In panel B, we see that both classic regression and a more flexible general additive model (GAM) both have the tendancy for overfitting to the most numerous sampled groups. The mixed effect model produces a fit that is close to the "truth" or data generating model.

.

4 Key notes for mixed-model selection

Each of the above mentioned approximation methods, have certain benefits and disadvantages. I list some considerations below, for each method:

1. PQL is fast, yet yields biased estimated when variances in random effects (the sd's) are large. It is especially biased with binomial data or when N per effect is low (e.g. ¡ 5 per random block). PQL also gives quasi-

likelihood which many statisticians feel cannot be used in inference (Wald, Z and T statistics) or selection (e.g. AIC, DIC). Basically all inferences based on the likelihood are invalid in combination with PQL (Joe, 2008). P-values and CI are complicated to calculate [1].

2. LA, is less biased, and approaches the real ML and can therefore be used in likelihood based inference test - however it assumes the likelihood distribution is approximately normal. It is also slower and less flexible than PQL. P-values and CI are complicated as in PQL.

3. GHQ, are even more precise than LA and also approach the real ML. It is much slower than LA and fitting models with more than 2–3 random effects is not considered feasible. P-values and CI are complicated as in PQL.

4. MCMC methods are highly flexible, can handle many random-effects, and are theoretically well founded with the "Bayasian Framework". CI intervals and "p-values" (i.e. quantiles) are simple to calculate from the posterior distribution samples. They are notoriously slow however, and technically challenging to implement. MCMC methods give very similar answers to the previous 3 methods when datasets are informative and priors weak.

5 Problems associated with Random effects

Statistics as Z, t, χ^2 and F are poor for models containing random effects as standard deviations are strictly positive (≥ 0), and thus violate the null hypothesis assumption ($\sigma = 0$). Likelihood ratio tests are also problematic and highly unsuited for PQL estimates. Although when using LA and GHQ, likelihood ratio tests can be used on random effects in some cases, however not if using REML. Guidelines on these issues are sparse, and I will update this section if more information arises.

Another considerable problem in model selection and inference, is how to decide how many parameters a models has (in AIC) or df (in t, χ^2 and F) when including random effects. How many parameters to you effectively have? With random effects included, there is no straight answer (Grueber *et al.*, 2011) and note that simply counting the random effect sd's is considered wrong. I have found no satisfactory answer to this, barring the use of DIC with MCMCs. DIC uses pD ('the effective number of parameters') instead of the number of parameters (Spiegelhalter *et al.*, 2002). The idea behind pD is that it is a more appropriate measure of model complexity than parameters alone which may say little of how complex a model is to fit.

One strategy that can be used in combination with model selection tools as AIC in combination with random effects, is to select among model with the same

[1] see Douglas Bates rant on the matter: https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

random effects fit to the same data. As the random effects are equal between models, ranking will depend on the approximated likelihood and fixed effects (this is only valid for LA and GHQ). However, this still leaves the question on how to select among models with different random effects open.

One suggestion, for model selection between models with different random effects, is to use a likelihood ratio test (Pinheiro & Bates, 2000, pp. 83-87). However, also here some issues arise, mostly from the fact that the parameter value under the null hypothesis is on the border of the parameter space. Technically, there is a possible correction, applied to the degrees of freedom of the LRT statistic, which partly takes account of this. In short, it uses an equal mixture of chi-squares with 2 different degrees of freedom. However, Pinheiro & Bates (2000) note, as verified by simulation, that this commonly suggested correction is not really quite right either. They conclude that it is tricky to do this just right, no easily-implemented method seems to fix the problem exactly, but the simple, naive LRT - being slightly conservative - is the recommended approach.

6 Some Important Mixed Model Assumptions

1. *Assumption of normality of random effects.* In mixed effect models one assumes that random slopes or intercepts come from a single probability distribution and we estimate the parameters of that distribution rather than (formally) estimating the individual intercepts of e.g. a randomized block design. The usual assumption is that the distribution of the intercepts is normal, though this can be relaxed in Bayesian models for instance. Thus a typical random (intercept) effects model will look like this;

$$\begin{aligned} Y_{ij} &= \beta_0 + \mu_{0i} + B_i X_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &\sim Normal(0, \sigma^2) \\ \mu_{0i} &\sim Normal(0, \tau^2) \end{aligned} \tag{6.1}$$

   Here i denotes the "random block", for which random intercepts are estimated and j the individual observations in that block i. Observations at level j are usually assumed independent while observations over level i usually not. We see that both μ_{0i} and ϵ_{ij} are assumed assumed to be are independent and normally distributed with mean 0 and variance τ^2 and σ^2 respectively. And that residuals ϵ_{ij} are calculated taking the random intercepts into account! In a GLMM the normality assumption of the errors ϵ is relaxed, however the normal assumption on the random intercepts remains. The normality assumption of random effects can be further relaxed in a Bayesian framework.

2.

7 The P-value controversy

P-values are a tricky issues with mixed-models. Confusion reigns [2] on how the p-values should be computed. And because of this confusion, Doug Bates declines to provide p-values in lme4 [3]. I will try to summize the probelms, gathered from various online sources, shortly here.

P-values are in the traditional sense are a test of a null hypothesis. They are based on computing the probability of observing a test statistic as extreme or more extreme than the observed, if the null hypothesis is true. By convention, we reject the null hypothesis if the p-value is less than some threshold (often 0.05). P-values can be calculated once the "sampling distribution" or the asymptotic distribution of the test statistic under repeated sampling with equal sample size is known. We then use the cumulative distribution function of this distribution to calculate p-values (under the assumtion that null hypothesis is true). Probably the best know example of this is the t-distribution, used when we compare e.g. two normal means (given they have equal variances or adjusting the degrees of freedom when they have uneuqla variances and so on).

With mixed-models, the cumulative distribution function of the test statistic when the null hypothesis is true is simply not known (as is the case in many other hierarchical models). So, without the "sampling distribution" to compute the p-value, what do we do? It turns out that for a limited range of hierarchical models e.g. split-plot designs in analysis of variance the reference distribution is known (it's called the F). However the rules for these cases do not necessarily translate to the analysis of any arbitrary hierarchical design, which might be unbalanced, and have crossed and correlated random effects.

Even if we can assume that the distribution in these more complex cases is the F, we still need to define the degrees of freedom. The numerator degrees of freedom are obvious (variance between models), but the denominator degrees of freedom are not so easily calculated (see also the problems accociated with AIC above). Numerous ways on how to adjust the denominator degrees of freedom have been suggested but this still brings uss back to the first problem that it is anything but clear if the reference distribution is truely F, and therefore it remains a question if correcting the denominator degrees of freedom solves anything.

8 Please contribute

Any comments or suggestions on how to improve this document, and make it a more comprehensive guide to inference with Mixed Effect models, are welcome. You can add your suggestions or comments through [4] or fork this repository on github [5].

[2] https://stat.ethz.ch/pipermail/r-sig-mixed-models/2008q2/000904.html
[3] https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html
[4] `http://github.com/MarcoDVisser/mmnotes/issues`
[5] `http://github.com/MarcoDVisser/mmnotes`

References

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Gelman, A. (2006) Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, **48**.

Grueber, C.E., Nakagawa, S., Laws, R.J. & Jamieson, I.G. (2011) Multimodel inference in ecology and evolution: challenges and solutionsn. *Evolutionary Biology*.

Joe, H. (2008) Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, **52**, 5066–5074.

Pfister, C.A. & Stevens, F.R. (2003) Individual variation and environmental stochasticity: implications for matrix model predictions. *Ecology*, **84**, 496–510.

Pinheiro, J.C. & Bates, D.M. (2000) *Linear mixed-effects models: basic concepts and examples*. Springer.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.