Mixed Models are Sometimes Terrible

Christopher Eager & Joseph Roy

Department of Spanish & Portuguese, School of Literatures, Cultures & Linguistics, University of Illinois at Urbana-Champaign

Introduction

Mixed-effects models have emerged as the “gold standard” of statistical analysis in different sub-fields of linguistics (Baayen, et al., 2008 ; Johnson, 2009; Barr, et al, 2013; Gries, 2015). One problematic feature of these models is their **failure to converge** under maximal (or even near-maximal) random effects structures. Convergence tests are themselves different from version to version of statistical packages for mem and also differ across platforms (e.g. R, SPSS and SAS). The lack of convergence is relatively unaddressed in linguistics and when it is addressed has resulted in ad-hoc statistical practices (e.g. Gries, 2015; Bates, et al, 2015; Jaeger, 2009) that are not found in the statistical literature on mixed models (e.g. Demidenko, 2013) and are premised on the idea that non-convergence is an indication that a random effects structure is over-specified (or not parsimonious).

Parsimonious Convergence Hypothesis (PCH)

The Parsimonious Convergence Hypothesis motivating these approaches is that **the failure of mixed effects model to converge is (most likely) due to the incorrect specification of the random effect structure**. In it is not clear from the statistical or applied statistical literature that convergence and parsimony are linked. Following the advice of Hodges (2014) to *pry open the black box* of mixed effects models, this study experimentally tests the PCH with near-balanced (simple) data and moderately to severely imbalanced (complex) data.

Ad-hoc Practices

When a Mixed Effects Model doesn’t converge

Under the PCH, failure to converge is taken to be a mis-specification of the random effects structure. The desire to fit maximal models have resulted in **ad-hoc practices**, some **reasonable*** (but unattested in the formal statistical literature on these models) and some clearly **unreasonable**.

1. Use a PCA of covariance matrix to determine most meaningful slopes (Bates, et al., 2016).
2. Reduce item random effect structure then reduce subject random effect structure until convergence (Jaeger, 2009).
3. Start with intercept only, use anova() to determine if a slope should be added or not. Stop when all slopes are not significant (Gries, 2015).
4. Keep removing slopes randomly from item and subject random effect structure until convergence.
5. Suppress or ignore convergence errors.

* Reasonable approaches can still lead to incorrect conclusions in some circumstances as we show for (1).

Method

Simulation Data Sets Types

Simple Linear and Logistic: With the simulation assumptions reported in Barr, et al. (2013) with 24 subjects and 24 Items and one binary predictor and removing up to 5% of the data.

Complex Linear and Logistic: With moderate and severe levels of imbalance in the data, one binary and one three level predictor with a true, non-zero maximal random subject effect structure. The number of subjects vary between 30 and 60. The mean number of observations per subject varies between 20 and 30.

Balance: A balance ratio (from 0, indicating total imbalance to 1, indicating complete balance) was used to measure the amount of imbalance within-subject for the complex data sets.

Fully Specified Bayesian Model

$$\beta_j \sim N(0,2)\ for\ j \in \{0,1,2,3\},$$

$$\sigma_{\mathcal{S}j} \sim HalfN(0,1)\ for\ j \in \{0,1,2,3\}, \qquad \Omega_{\mathcal{S}} \sim LKJ(2),$$

$$\Sigma_{\mathcal{S}} = \mathrm{diag}(\sigma_{\mathcal{S}})\Omega_{\mathcal{S}}\mathrm{diag}(\sigma_{\mathcal{S}})^T, \qquad \gamma_{\mathcal{S}} \sim MN(\mathbf{0},\Sigma_{\mathcal{S}}),$$

$$[Linear] \quad \sigma_\epsilon \sim HalfN\left(0,\frac{1}{2}\right), \qquad y \sim N(X\beta + Z\gamma,\sigma_\epsilon),$$

$$[Logistic] \quad \ln\left(\frac{p}{1-p}\right) = X\beta + Z\gamma\,, \qquad y \sim Bernoulli(p).$$

Following Kimball, Shantz, Eager & Roy (2016) and implementable with Eager (2017), the above weakly informative constraints were used in order to estimate the mixed effects parameters in RStan (Carpenter, et al. In press).

Convergence

Tests of convergence are meant to assess whether the estimates for a mixed effects model are statistically and computationally reliable.

We take convergence in **lme4** to mean **no errors or warnings have been produced while setting the tolerance to .01** (lme4 default is .002).

In **RStan**, we take convergence to mean **no divergent transitions and all R-hats less than 1.1**.

Results

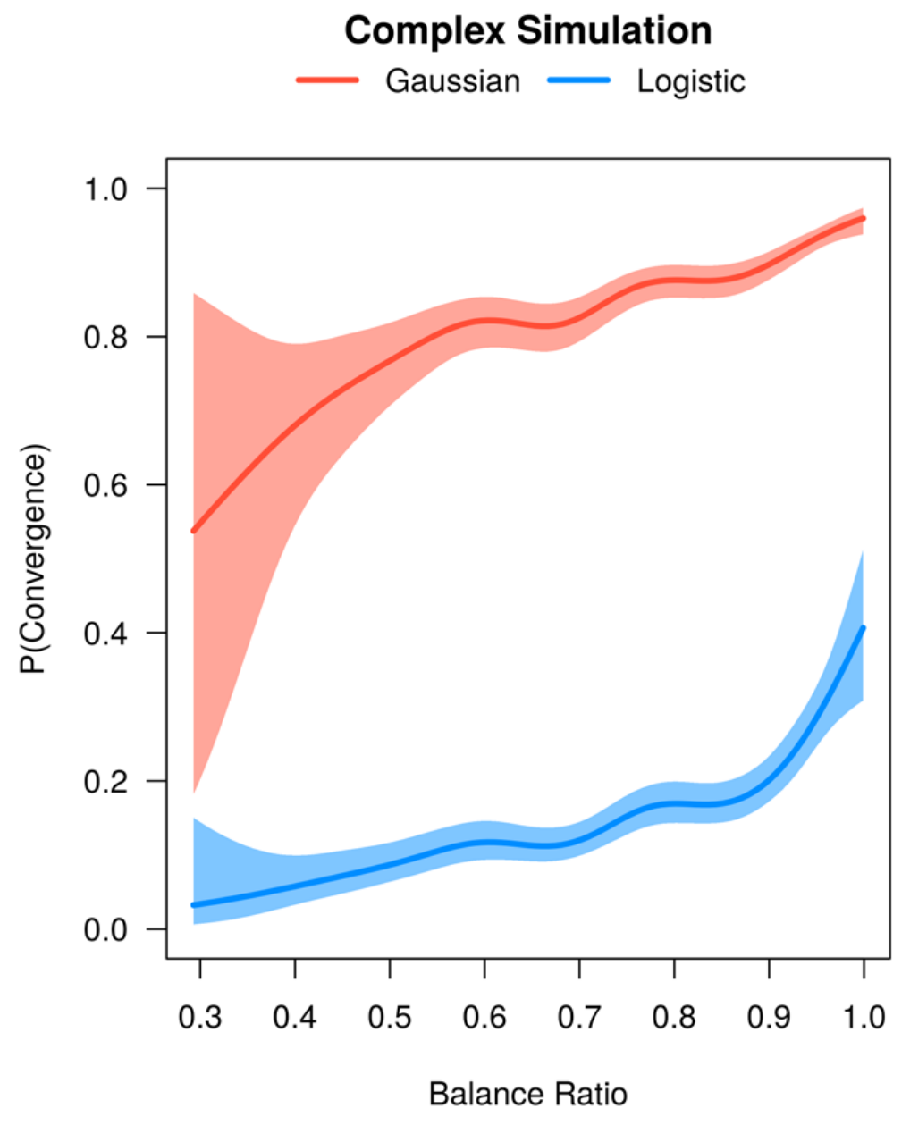
Type	lme4	RStan	# of Sims
Simple Linear	0 %	.009 %	80,000
Simple Logistic	7 %	.002 %	20,000
Complex Linear	14 %	3 %	2,500
Complex Logistic	82 %	<.001 %	2,500

Table 1: Rates of non-convergence in unconstrained mixed effects (lme4) and fully specified Bayesian models (Rstan)

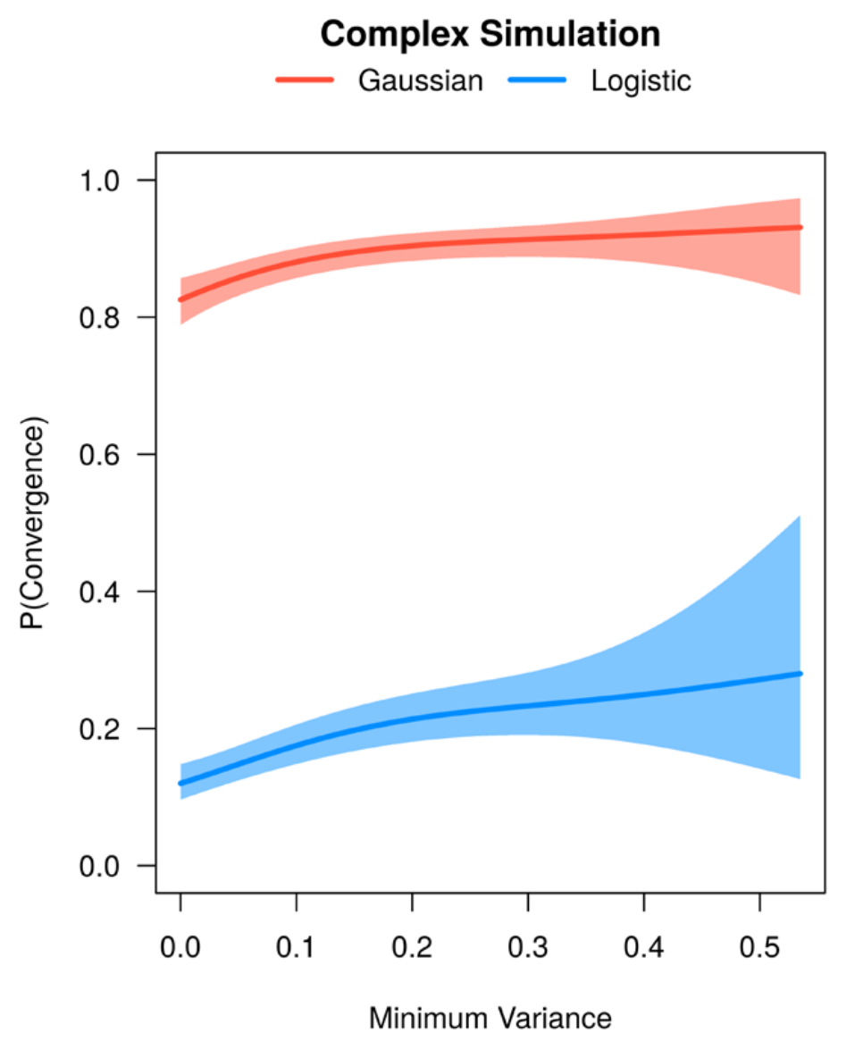
PCA on Correlation Matrix

For the complex converged models, rePCA(), correctly identified the random effects structure in **17** out of 445 (3%) logistic mixed models and **985** out of 2148 (45%) linear mixed models.

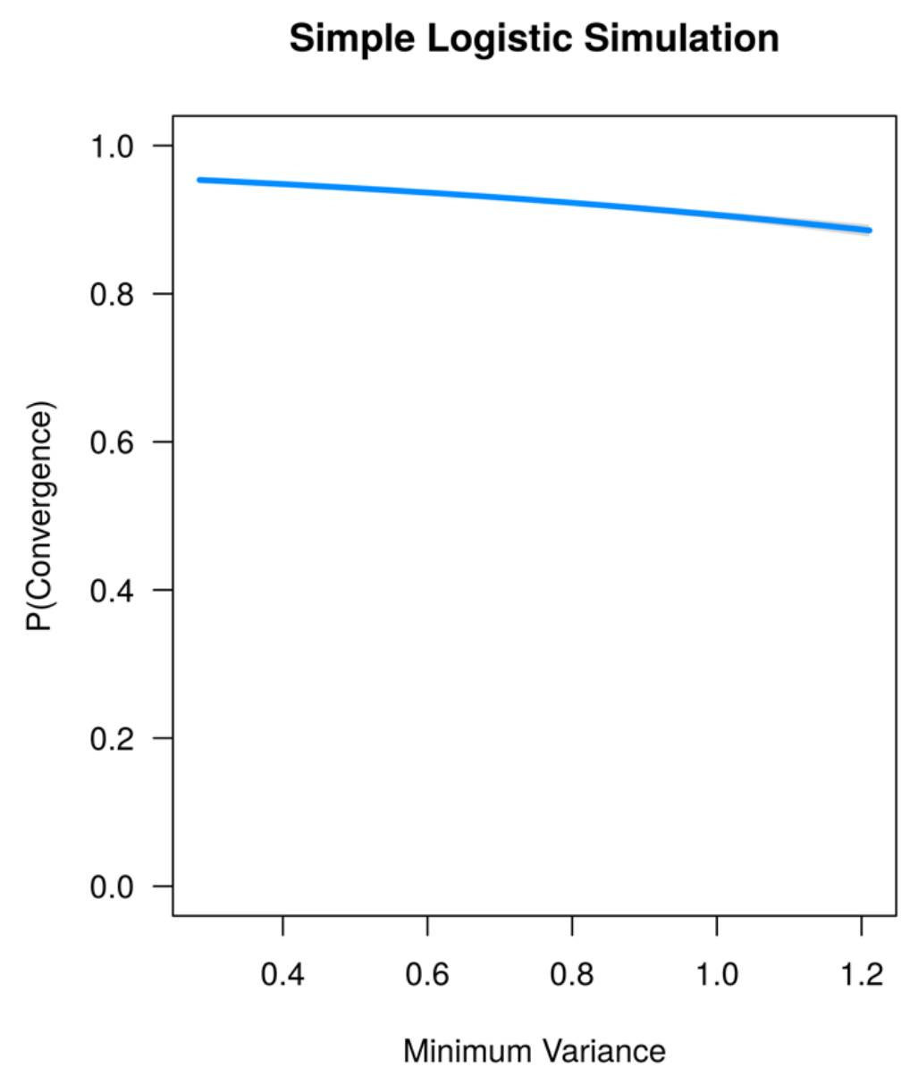
Balance and Minimum Variance



Balance has the expected effect on convergence in both logistic and linear mixed models. More balance implies more convergence for the correct model.



Examining the effect of minimum random intercept and slope variance in the model, an increase in variation seems to increase the likelihood of convergence.

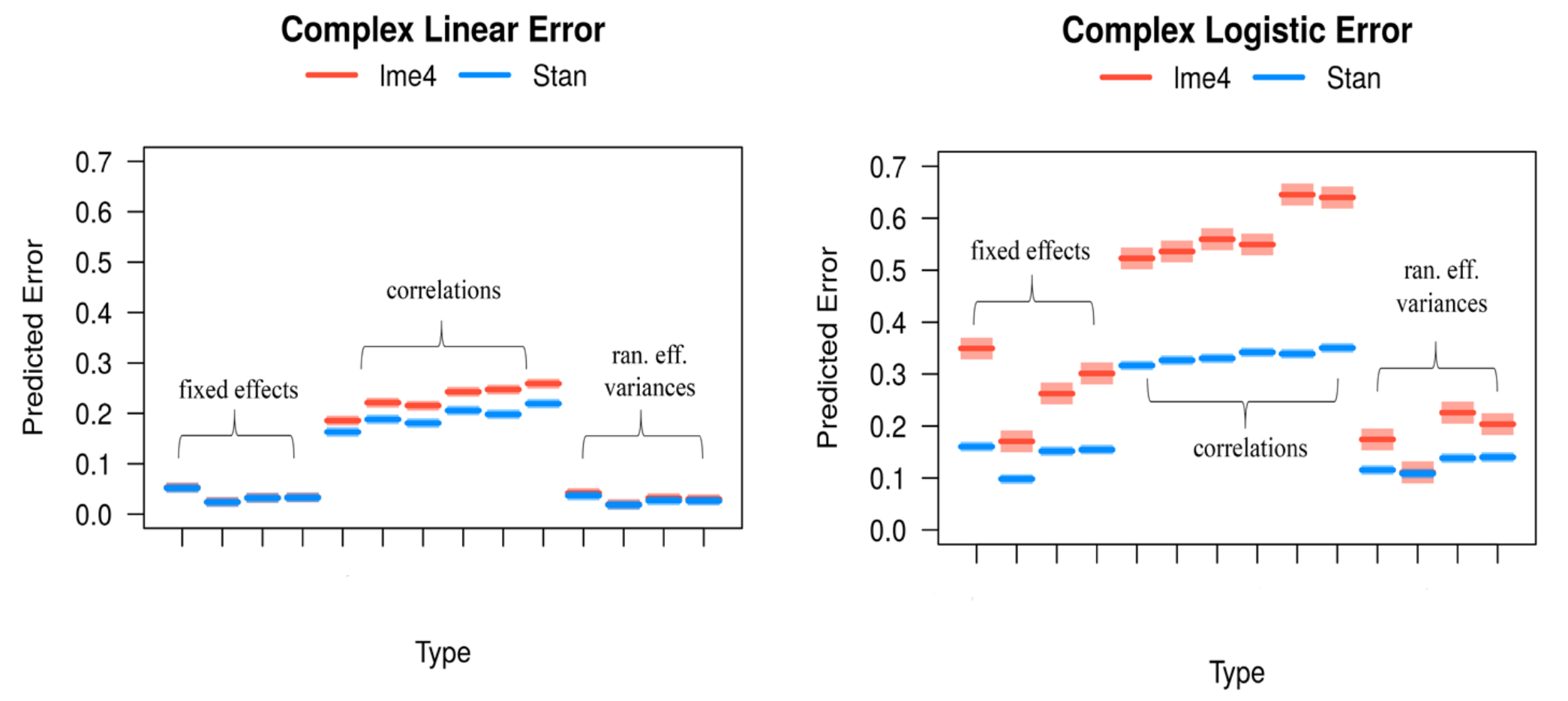


In the simple logistic model, the minimum random effect variance has an interesting effect in that likelihood of convergence decreases as the minimum variance increases.

This may indicate it is more difficult to locate a solution for the fixed effects and random effects when there is more noise in the data with near balance in the design.

Converged Model Predicted Error

Even when the complex models converged, did RStan or lme4 do better at estimating parameters?



For linear regression, there is a small difference in error for correlations, but both random effect variance and fixed effects show no difference between the two software types. For logistic regression, however, Stan does better at estimating all model parameters.

Conclusions

The terribleness alluded to in the title refers not to what these models are intended to do in linguistics (i.e. account for multiple observations on same participant or item), but instead how these models behave with imbalanced logistic regressions as well as the PCH which linguistic researchers have used to guide their behavior when confronted with model non-convergence. This project has shown, with both real data (Kimball, Shantz, Eager and Roy, 2016) as well as the simulations discussed in this poster, that this terribleness can be alleviated with more constraints than standard mixed effects models implemented in lme4 provide: namely, by using a more fully specified Bayesian model (Eager, 2017). Further, when there is convergence, RStan does better for logistic models in estimating fixed effects and random effects in our simulations for moderate to severely imbalanced data.

Next Steps

1. Simulate models with known zero random effects.
2. Continuous imbalance (i.e. coverage)
3. Power: a wider range of subjects/observations per subject.

Selected References

Carpenter, Bob, et al. In press. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*

Eager, Christopher. 2017. *Stanmer* https://github.com/CDEager/stanmer

Kimball, Shantz, Eager & Roy, 2016. Beyond Maximal Random Effects: Moving past convergence problems. https://arxiv.org/abs/1611.00083

Hodges, James. 2014. *Richly Parameterized Linear Models: Additive, Times Series, and Spatial Models Using Random Effects*. Boca Raton, FL: CRC Press

Demidenko, Eugene. 2013. *Mixed Models: Theory and Applications with R*. 2nd Ed. Hoboken, NJ: Wiley.

Acknowledgments

The Office of the Vice Chancellor for Research and the School of Literatures, Cultures and Linguistics at the University of Illinois at Urbana-Champaign provided funding for this research.

ILLINOIS