Do Bayes Factors and mixed effects models mix?

Lauren Kennedy

Econometrics and Business Statistics, Monash University, Australia

Abstract

While Bayes Factors have become common practice in psychology, questions remain about how to use them. In response to the target article, I argue that one of the reasons that Bayes Factors are challenging to implement with mixed effects models is that they force the user into a difficult interpretation of random effects. In this article I discuss two reasons for this. Firstly the null hypothesis significance test framing required to make dichotomous model judgements, and secondly the strong divide between explanatory and predictive methods of model comparison in psychology.

Do Bayes Factors and mixed effects models mix?

Introduction

The target article by van Doorn, Aust, Haaf, Stefan, and Wagenmakers (2021) describes the use of Bayes Factors when modelling repeated observations using mixed effects models. They note three practical challenges in their article, which stem from taking a binary, explanatory based approach to data. In this comment I discuss some alternative approaches to this conceptualisation of the problem, notably the choice of explanatory vs predictive frameworks and the binary null hypothesis significance testing approach discussed in the target article.

Although the target article is specifically focused in Bayes Factors as applied to multilevel models, I suggest in this comment that Bayes Factors, which arguably fall in the ‘explanatory‘ family of model prediction[1], might not be the most suitable tool for multilevel models at all.

Null hypothesis significance testing

Psychology has long since marked itself as a field that uses null hypothesis significance testing. Style guides taught students to prepare a block of test information (commonly p-value, degrees of freedom, and test statistic) (Burton, 2010) when discussing result significance. While the p-value has fallen out of favour, it appears to have been replaced with a Bayes Factor as a tool for understanding the impact of a particular parameter of interest. In this case, both p-values and Bayes Factors represent the comparison between a target hypothesised model and a null model, which is generally nested in the more complex model.

This means, of course, that a null model must be specified. The complexity of this problem means that much of the target article concerns itself not with how to calculate Bayes Factors nor what they mean for a multilevel model, but rather with which null model the alternative model should be compared to. However, binary model comparisons

[1] In their simplest form, a Bayes Factor is the ratio of the probability of data given hypothesis 1 over the probability of data given hypothesis 2 (Kass, 1993), although more complex interpretations are possible.

are not the only approach to model comparison. In the rest of this article, I demonstrate how predictive accuracy allows multiple models to be compared simultaneously.

In the target article, the null model needs to be specified to allow the researcher to correctly identify whether an effect of condition is present. The null models in consideration are models 3 (no condition effect, random intercept for participants) and 5 (random intercept for participants, random slope centred at 0 for condition). Interestingly, the alternative model also changes depending on the format of the data. The two alternative models considered are model 4 (random intercept for participant, condition effect) and model 6 (random intercept for participants, random slope with mean v for condition). Distilling this down, the model comparison techniques appear to differ predominately in their treatment of between-participant variation in the effect of condition (the presence of a random slope) or not.

However, the models available for comparison are not a full cross-section of potential models. They exclude models where the participant slope and intercept are correlated, where there is a random slope but not a random intercept, and models where a fixed effects approach is used to control for participant level variation. Considering the set of potential models, it seems apparent that the central question is whether there is an effect of condition or not, with some uncertainty over whether a random slope should be included.

What purpose do random effects serve?

Consider a simple example of individuals undertaking a task multiple times over, with one of two conditions assigned for each task. As this is a repeated measures design, we would be encouraged to fit a random effect for participant intercept and perhaps participant slope. However, the interpretation of why these random effects are used differs based on the interpretation of the theoretical framing of modelling. To explore why this is, we first consider random effects in the case of a prediction motivated framing, and then in terms of a theoretically motivated framing.

To understand the prediction motivated framing, consider that instead of a random effect of participant, we could instead dummy code participant ID and fit a single parameter for each and every participant (minus one for identifiability reasons). If there were many observations per participant, then we would see little difference between the parameter estimates in the multilevel model and the parameter estimates in the dummy-coded model. However, if there were only a few observations per participant, we would potentially see a larger difference. Notably, not only will the individual estimates be different in the second case, the uncertainty around these estimates will be different (usually smaller). Applying a random effect term involves a trade-off between bias in the participant-wise estimate, and a reduction in variance. Similarly, we could make this same decision when considering between dummy-coded, random or fixed effects for the effect of condition.

An alternative interpretation of the random effect is the conceptual framing a distribution of individual variation in the population. This is a theory-driven interpretation, but even theory in this context can be reframed as generalisation. Generalisation, or an interest in extrapolating beyond the observed sample, can easily be thought as prediction focused - prediction of those unobserved.

Even if a theoretically motivated interpretation can be decoupled from a statistically motivated example, in the target article, it seems that the random slope variation is not theoretically motivated at all. If it were, the argument for its inclusion in either the null or alternative model would surely have been theoretical grounds rather than statistical, and so discussion from a statistical framing would be unnecessary.

Another feature of using a random effect for the intercept and slope is that it not only provides greater predictive accuracy with respect to observed individuals, it also provides an avenue for for predicting new unobserved individuals. Varying components enable prediction for a new and unseen unit by drawing from the underlying distribution (in this case the intercept term for a new participant would be assumed to be drawn from a distribution given by the fitted random effect).

Prediction vs explanation

The target article does not consider why we need to compared different nested models in the first place. To do so, we need to take a step back and consider what it means for a model to be "good". For many in psychology, a "good" model is the one that best explains the observed data. Ideally, it will be the one that underlies the true data generating process or in the (almost certain) case that this is not possible, some close approximation of it. Null hypothesis significance testing is a useful methodology because to confirm that one model explains the observed data better than another model is to create evidence that one theory is more likely than other. Much of our work is concerned with this notion of theory, and the data we collect are a means of gathering evidence to either confirm or refute a theory.

This is not universally the case when discussing what makes a "good" model. For many in statistics, a "good" model is one that best predicts new observations of data from the same process. For example, in forecasting we might be concerned with predicting tomorrow's stock prices from what we have observed previously. This has led to the development of model comparison methods like leave-one-one-out cross-validation (LOO-CV) that aim to identify the best predictive model possible. However, these models often lack in interpretability — they are designed just to predict, not to understand. For more on the distinction between explanation and predictions, see Shmueli et al. (2010).

When we consider model comparison tools, we can categorise those that are designed to select models that predict, and those designed to explain. Prediction based methods focus on what is not observed and how well a model can predict it. Explanatory based methods focus on the observed data and how well we can explain it. Unsurprisingly we see model comparison tools that are popular in psychology (e.g., Bayes Factors) fall on the explanatory side, and model comparison tools that are popular in statistics (e.g., LOO-CV) fall on the prediction side.

It is understandable that there are strong opinions on either side — explanation ability does not predict predictive ability (Gronau & Wagenmakers, 2019a, 2019b; Vehtari, Simpson, Yao, & Gelman, 2018). However, I argue that there's a place for prediction based methods, particularly when comparing multilevel models. Although random effects models can be interpreted in a theoretical framework, in a prediction framework their potential for better predictive ability is demonstrated. Conversely, an explanatory approach with Bayes Factors does not consider the facets of predictive accuracy, even though predictive accuracy is one of the benefits of using multilevel models.

Although it's not clear why a random effect is used in the target article (particularly whether it stems from a theoretical or predictive framework), it seems evident that it is from a predictive framework. In particular, the target article employs a random intercept for participant across all of the comparison models (3-6), but the fuller version (fixed effect for each participant) is not considered. What purpose are the random component(s) serving? Are they a theoretically motivated question about individual differences, and if so, why is a random intercept approach considered instead of a fixed effects approach? Or are they used to provide stabilisation of individual estimates through partial pooling?

Prediction based methods for model comparison

An alternative to likelihood ratio-based methods (such as the BIC) are model comparison tools that evaluate the predictive ability of the model in question. These tools, such as the WAIC and LOO-CV have generally not been well received in psychology (Gronau & Wagenmakers, 2019a; Haaf, Klaassen, & Rouder, 2021). However, these explorations of predictive methods are generally based on simple and importantly non-regularising models. This means that the potential benefits of predictive methods that apply specifically when working with regularising models are not fully discussed.

Although I previously claimed the Bayes Factor is an example of a explanatory method, it can also be considered as a predictive method. As Gronau and Wagenmakers (2019a) note, a Bayes Factor can be considered a predictive method over the entire parameter space. However this means that the Bayes Factor loses one of the major benefits of predictive methods like cross-validation. This benefit is the ability to target the model validation technique specifically to the prediction that is most of interest in the particular application (e.g., see Burkner, Gabry, and Vehtari (2020)). Using the first case study of the target article, I explore this further.

One of the benefits of leave one out cross-validation is the intuitive extension to leave one component in the underlying structure of the data out and consider the ability of the model to predict for that component. For the first example in the target article, the data have a clear hierarchical structure — with random slopes and intercepts considered for participants. This structure (and more complex versions) is typical of all multilevel model designs. The expected log predictive density for leave one out cross-validation is

$$\mathtt{elpd}_{\mathtt{loo}} = \sum_{i=1}^{n} \mathtt{log}\ p(y_i|y_{-i}), \tag{1}$$

where the i^{th} data point is predicted given the remainder of the observed data. However, this may not be the prediction task that we are interested in. Instead we could consider the implicit structure of a hierarchical model to implement k-fold cross-validation, where the fold represents a particular level of a random effect. The benefit of this approach is that it both respects the implicit structure of the data (which neither Bayes Factors, LOO-CV or WAIC do), and provides for each model a number of metrics, all in terms of which aspect of the data is best fit by the data (is it the next trial block? a new participant? an additional condition?). In the context of the the first case presented in the target article there are three potential avenues for sensible folds.

The first is leave one participant out — where we leave each participant out in turn, refit the model and then predict this participant's responses. The second option is to fit the model for $t-1$ items using the full dataset, and then predict the remaining t^{th} item. This can be done if the ordering of the items is important (Burkner et al., 2020), but in this case the ordering of the items does not seem particularly important. Although not present in this data, a third option, if there were more than two conditions, is to leave out a condition and attempt to predict it.

Notably, as implemented in the LOO package (Vehtari et al., 2020), with standard error adaptions taken from Vehtari (2020), we obtain an estimate of not just the size of the expected log predictive density but the standard error. This allows us to compare the models but also understand the uncertainty of the model comparison metrics. While Bayes Factors give a sense of the size of the difference between models, conclusions from them can still be sensitive to type 1 and type 2 errors (Smithson, 2019).

In table 1 we present the expected log predictive density (ELPD) for each of the four models under consideration using the full data only [2] that although there is some movement within the full ordering of models, this is within the standard error for the

[2] Model implementation was conducted with rstanarm Goodrich, Gabry, Ali, and Brilleman (2020) using default priors. The aggregated data caused significant sampling issues so it is not presented

ELPD, and (unlike the Bayes Factor interpretation), model 6 remains the overall preferred model, but only by a marginal amount. When we consider leaving out single data points, the differences between ELPD are much greater relative to the standard error. This provides significant guidance as to where and when we expect our model to accurately generalise out of the sample, and when we would not.

Similar to the target article, we find that different specifications of our model comparison task can result in differences in the preferred model. Using a leave-one individual out procedure, we see that while there is some small evidence that model 6 is preferred over models 3 and 5, the standard error relative to these estimates is quite large. When we consider leave-one observation out, we see that the differences between model 6 and models 3 and 4 is substantially larger, but still with negligible differences between 5 and 6. From this we can surmise that there are few differences between the models regarding predicting for a new participant (although 6 is slightly preferred), but in regard to predicting the response for an observed individual, models with a random slope (5 and 6) are much more predictive than models without (3 and 4). Although the general trend is not substantially different from that observed with a Bayes Factor approach with full data, this framing allows greater specificity about what the models are differentiated on (new observations) and what there is limited differentiation on (new participants).

Prediction is not explanation

One of the major criticisms of prediction-based model comparison methods is that they do not not necessarily prefer the true generative model. Discussions around this challenge centre around M-open and M-closed theoretical frameworks. In a sense, this confuses the issue. Yes prediction-based model comparison techniques will not always favour the true generative model. However, these methods give us a sense of the difference between the variance explained in the data we have observed, and the variance explained in the future data we would hope to observe. The first type of variance is what is captured with a Bayes Factor (or AIC), and is typically used for theoretical advances. These methods help us to answer questions like "Is it more likely that a model with this variable included is closer to the true data generation process". If theory development is the pursuit of finding the true underlying model that produced the observed data, then considering the explanatory ability is important.

However, predictive models generally represent an upper bound of what the model could predict. The difference between explanation and predictive ability could represent avenues for better measurement and/or other methodological advances (Shmueli et al., 2010). In addition, prediction-based methods help in cases where models exhibit some or complete model mimicry, which is a serious problem for the complex models often used in mathematical psychology.

Lastly, I do wish to question the prevalent notion that predictive ability is not helpful for theory development. Presumably, the purpose of theory is to predict the behaviour of new individuals in new conditions or different situations, If we only select the model that has maximum explanatory ability on the data observed, what does this imply when future studies with new data/conditions fail to prefer the same model?

Priors and prediction

In the target article (van Doorn et al., 2021) note two implications of Bayes Factors in case 2. Firstly that Bayes Factors appear sensitive to the prior specification, particularly in terms of the random effect. Secondly, that there remains a statistical question of differentiating between models where there is at least partial model mimicry. By considering the predictive framework, the change in problem framing simplifies both questions.

The question of priors, particularly the priors of a varying effect, relies heavily on the notion of what priors contribute in a varying effects model. In a theoretical or explanatory view, priors are used to represent the previous knowledge or experience of the parameters in question. In case 2 of the target article, the priors represent a previous understanding of the size of the effect and the size of the measurement error.

However, priors mean different things when we consider a predictive approach. In a predictive approach, priors are used to induce some degree of regularisation or shrinkage. They might be motivated by previous data or experience, or motivated by prior predictive checks (Kennedy, Simpson, & Gelman, 2019), but the expectation is that priors will have an impact on model estimates. The purpose of their use in this instance is to accept some bias in model parameters to obtain better predictive ability of future data.

Conclusion

In the target article, van Doorn et al. (2021) propose a variety of alternative framing of Null Hypothesis Significance Testing with multilevel models. In response to this, I argue that multilevel models are particularly well suited to a predictive rather than an explanatory framework, and that this predictive framework negates the need to focus on binary model comparisons. I hope this aids the conversation of model comparison and selection with multilevel models, and in particular emphasises the potential for prediction based framing to complement our theory driven model development.

References

Burkner, P.-C., Gabry, J., & Vehtari, A. (2020, Jun). Approximate leave-future-out cross-validation for bayesian time series models. *Journal of Statistical Computation and Simulation*, *90*(14), 2499–2523. Retrieved from `http://dx.doi.org/10.1080/00949655.2020.1783262` doi: 10.1080/00949655.2020.1783262

Burton, L. J. (2010). *An interactive approach to writing essays and research reports in psychology.* John Wiley & Sons Australia, Ltd.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `https://mc-stan.org/rstanarm` (R package version 2.21.1)

Gronau, Q. F., & Wagenmakers, E.-J. (2019a). Limitations of bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior*, *2*(1), 1–11.

Gronau, Q. F., & Wagenmakers, E.-J. (2019b). Rejoinder: More limitations of bayesian leave-one-out cross-validation. *Computational brain & behavior*, *2*(1), 35–47.

Haaf, J. M., Klaassen, F., & Rouder, J. N. (2021, Mar). *Bayes factor vs. posterior-predictive model assessment: Insights from ordinal constraints.* PsyArXiv. Retrieved from `psyarxiv.com/e6g9d` doi: 10.31234/osf.io/e6g9d

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *42*(5), 551–560.

Kennedy, L., Simpson, D., & Gelman, A. (2019). The experiment is just as important as the likelihood in understanding the prior: A cautionary note on robust cognitive modeling. *Computational Brain & Behavior*, *2*(3), 210–217.

Shmueli, G., et al. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Smithson, M. (2019). *Can bayes factors "prove" the null hypothesis?*

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A., & Wagenmakers, E.-J. (2021, Feb). *Bayes factors for mixed models.* PsyArXiv. Retrieved from `psyarxiv.com/y65h8` doi: 10.31234/osf.io/y65h8

Vehtari, A. (2020). *Cross-validation for hierarchical models.* Retrieved from `https://avhetari.github.io/modelselection`

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.* Retrieved from `https://mc-stan.org/loo/` (R package version 2.4.1)

Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2018). *Limitations of "limitations of bayesian leave-one-out cross-validation for model selection".*

Model	LOO technique	ELPD diff (vs model 6)	SE diff (vs model 6)
Model 3	leave one ind out	-4.0	3.9
Model 4	leave one ind out	-1.2	3.6
Model 5	leave one ind out	-3.3	2.9
Model 3	leave one obs out	-85.0	12.9
Model 4	leave one obs out	-72.4	11.6
Model 5	leave one obs out	-0.3	0.5

Table 1

Leave one individual out and leave one observation methods comparing the four models introduced in the target model. Negative values indicate worse prediction when compared with the comparison model (model 6 in all cases).