Statistical Science
2007, Vol. 22, No. 3, 349–352
DOI: 10.1214/07-STS235A
Main article DOI: 10.1214/07-STS235

Comment: Bayesian Checking of the Second Levels of Hierarchical Models

Andrew Gelman

Bayarri and Castellanos (BC) have written an interesting paper discussing two forms of posterior model check, one based on cross-validation and one based on replication of new groups in a hierarchical model. We think both these checks are good ideas and can become even more effective when understood in the context of posterior predictive checking. For the purpose of discussion, however, it is most interesting to focus on the areas where we disagree with BC:

1. We have a different view of model checking. Rather than setting the goal of having a fixed probability of rejecting a true model and a high probability of rejecting a false model, we recognize ahead of time that our model is wrong and view model checking as a way to explore and understand differences between model and data.
2. BC focus on p-values and scalar test statistics. We favor graphical summaries of multivariate test summaries.
3. For BC, it is important that p-values have a uniform distribution (i.e., that they be u-values, in our terminology) under the assumption that the null hypothesis is true. For us, it is important that p-values be interpretable as posterior probabilities comparing replicated to observed data.
4. BC recommend an "empirical Bayes prior p-value" as being better than the posterior predictive p-value. In fact, their empirical Bayes prior p-value is an approximation to a posterior predictive p-value which was recommended for hierarchical models in Gelman, Meng and Stern (1996). BC miss this connection by not seeing the full generality of posterior predictive checking.

In our discussion, we go through each of the above points in turn and conclude with a comment on the potential importance of theoretical work such as BC's on the future development of predictive model checking.

Andrew Gelman is Professor, Departments of Statistics and Political Science, Columbia University, New York, New York 10027, USA (e-mail: gelman@stat.columbia.edu)

1. THE GOAL OF MODEL CHECKING: REJECTING FALSE MODELS, OR UNDERSTANDING WAYS IN WHICH THE MODEL DOES NOT FIT DATA

All models are wrong, and the purpose of model checking (as we see it) is not to reject a model but rather to understand the ways in which it does not fit the data. From a Bayesian point of view, the posterior distribution is what is being used to summarize inferences, so this is what we want to check. The key questions then become: (a) what aspects of the model should be checked; (b) what replications should we compare the data to; (c) how to visualize the model checks, which are typically highly multidimensional; (d) what to make of the results?

In a wide-ranging discussion of a range of different methods for Bayesian model checking, BC focus on the above question (d): in particular, how can Bayesian hypothesis testing be set up so that the resulting p-values can used as a model-rejection rule with specified Type I errors? This question is sometimes framed as a desire for calibration in p-values, but ultimately the desire for calibration is most clearly interpretable within a model-rejection framework. For example, BC write that some methods "can result in a severe conservatism incapable of detecting clearly inappropriate models." But it is not at all clear that, just because a model is wrong, that it is "inappropriate." If a model predicts replicated data that are just like the observed data in important ways, it may very well be appropriate for these purposes. Recall that we have already agreed that our models are wrong; we would like to measure appropriateness in a direct way, rather than set a rule that even a true model must be declared "inappropriate" 5% of the time. For example, in the model considered by BC, we do not see the rationale for their testing the hypothesis $\mu = \mu_0$; we would rather just perform Bayesian inference for μ.

Our concerns are thus a bit different from those of BC: we are less concerned about the properties of our procedures in the (relatively uninteresting) case that the model is true, and more interested in having the ability to address the misfit of model to data in direct terms.