Rejoinder

Andrew GELMAN

I appreciate Buja's generous comments and will briefly clarify some issues regarding the role of data visualization in model checking, and the relevance of Bayesian inference to model checking.

1. DATA VISUALIZATION

My article presents four general kinds of model-based graphical diagnostics:

1. Displays of raw data, with a model used to simulate the reference distribution of the displays (as in Figure 1). Conversely, Figure 2 illustrates the difficulty, in general, of interpreting data displays *without* a comparison to a reference distribution.
2. Similar displays of tests of lower-dimensional data summaries (e.g., scalars in Figure 3, and vectors in Figures 4 and 5(a)) compared to simulations from fitted model—again, with discrepancies from the simulations illustrating aspects of the data not explained by the model.
3. Residual plots—put more generally, graphs of differences between data and fitted model which, if the model is true, should follow distributions with invariance properties such as independence and zero mean. Figures 5(b), 6, and 7 illustrate how violations of these invariance properties are visibly apparent without the need for explicit comparisons to simulated replications.
4. Displays of latent data/parameters (as in Figures 8 and 9) or of completed datasets (i.e., combinations of observed and missing or latent data, as in Figure 10).

In methods 1 and 2 above, the displays and test summaries are functions of data alone, with the role of the model being to define a reference distribution for comparison. Thus, if a series of models is developed for a single dataset, one would hope to see the replications looking more and more like the actual data. An attractive example appears in Ripley (1988, chap. 6).

*Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 785–787
DOI: 10.1198/106186004X12317*

In methods 3 and 4, the model is used to construct the display, and as improved models are fit to a single dataset, the plots should look increasingly "reasonable" in the sense of being consistent with invariances in the model and with outside knowledge of what the completed dataset should look like.

In all these approaches, the question arises in application of how extreme are the departures from the model. Buja's visual permutation test and simultaneous inference bands are important ideas that go beyond simple display of reference distributions to more systematic or focused comparisons. The visual permutation test seems particularly promising for automatic computer implementations of predictive tests.

Data visualization ideas—from the classical methods of Tukey, Cleveland, and others, to the more recent dynamic approaches of Buja, Cook, and their collaborators—should be useful in understanding model fit by bringing data, residuals, and latent data closer to the user, and in allowing visual comparisons to ever-more-complicated reference distributions. We also anticipate that specific graphical displays in fields such as computer science, genetics, and social networks can be made more effective by explicitly displaying reference distributions alongside data displays, or by constructing plots of residuals or latent-data plots that would reveal structure over and above what is expected in fitted models.

On a more specific point, I thank Buja for pointing out the error in the claim that the Kolmogorov-Smirnoff test could be directly used to detect the misfit in our mixture model example. As he points out, performing this model check requires a more elaborate treatment of the replication distribution. Another possibility would be to use a directional discrepancy, $T(y, y^{\text{rep}}) = \sup_x (F_y(x) - F_{y^{\text{rep}}}(x))$, rather than the Kolmogorov-Smirnoff distance, $T^{\text{K-S}}(y, y^{\text{rep}}) = \sup_x |F_y(x) - F_{y^{\text{rep}}}(x)|$. Unlike $T^{\text{K-S}}$, the directional discrepancy T is antisymmetric in y and y^{rep} and has the property that, if the model is true, its distribution is symmetric about zero. However, for this example the directional discrepancy will not actually detect the model misfit, because it could be equally likely to be positive or negative here. So Buja is correct that more effort would be needed to numerically summarize this model misfit that is visually so clear. The suggestions in his discussion provide some interesting directions for general comparisons of distributions.

2. BAYESIAN DATA ANALYSIS AND GENERATIVE MODELS

Buja questions why I focus on Bayesian posterior predictive distributions to the near exclusion of other approaches for creating reference distributions such as permutation tests, bootstraps, and cross-validation. The immediate motivation is probably from seeing so many Bayesian analyses with the following pattern:

1. "Exploratory data analysis": simple displays of raw data; for example, histograms or scatterplot matrices.
2. Construction of a series of more complicated models.
3. Extensive discussion, often including graphical diagnostics, of the convergence of iterative simulations.

4. Presentation of parameter estimates and uncertainties (typically in simple tabular form), to conclude the analysis.

These analyses rarely feature model checking or model-based graphical displays. Sometimes there is model comparison or model averaging, featuring numerical measures such as BIC or DIC, but rarely a graphical check showing the implications of the entire fitted model.

In some sense, the rarity of model checks can be understood on sociological grounds: if the final model does *not* fit the data, researchers have an obligation to improve the model until it fits, at which point a diagnostic check would be unnecessary. However, we suspect that the real issue is that models are checked little if at all.

I believe that one reason for Bayesian models not being checked is the perception that model checking is not necessary or even appropriate in Bayesian inference—an attitude we associate with a superficial reading of Savage, Lindley, and other subjective Bayesians. In our recent research we have tried to provide a theoretical foundation and examples of Bayesian predictive model checking (see Gelman et al. 2003, chap. 6).

In short, this article focused on Bayesian model checking partly because of our own positive experiences with Bayesian inference, but also because it seemed to us that Bayesians had the most to gain from model checking. Non-Bayesians seem more aware of the potential problems of using wrong models, and we wanted to show Bayesians how effective exploratory data analysis can be used, if taken seriously as part of the iterative model-checking process.

More generally, our approach requires a *generative model*—that is, a probability model that has the potential to generate observed data along with hypothetical replications. As noted by Buja, posterior inference has the additional advantage of automatically generating a set of generative models, thus separating inferential from predictive uncertainty—but for most applications, we have found the predictive uncertainty to be the most important, which is why we suspect these approaches could be nearly as successful with maximum likelihood or bootstrap inference as with full Bayes.

However, we would anticipate more difficulty applying EDA approach to statistical methods that are not fully model-based. Here we are thinking of methods such as quasi-likelihood, generalized estimating equations, and probability-weighted estimates for censored data, which produce parameter estimates without explicitly modeling the data-generation process. (For example, marginal models in biostatistics for longitudinal data estimate unit-level regression coefficients without fully specifying a model for the observations at each time point.) For these methods, it is not so easy to generate hypothetical replications, or random imputations of missing or latent data, and so plots of type 1, 2, and 4 (see the beginning of this rejoinder) cannot be routinely implemented. This is one reason we prefer to use generative models, whether or not their parameters are estimated Bayesianly.