Bayesian Analysis (2010) **5**, Number 4, pp. 691–696

Comment on Article by Vernon et al.

David A. van Dyk*

This paper tackles the computationally challenging task of comparing the predictions of the sophisticated Galform computer model for Galaxy Formation to observed light curves—data on the number of galaxies observed per unit volume in a given bin of luminosity for a particular band of light. The authors are to be commended for their clearly careful and diligent model-checking of this complex computer model. Judging from Figures 12 and 13 they where able to find parameter values that agree much more closely with the observed luminosity functions then what was previously available. (Although when comparing with data for which the model was not tuned, as in Figure 14, the results are more ambiguous.) By exploring the distribution of the parameters that result in acceptable model fits, the authors are able to draw conclusions about the complex relationships among the parameters of scientific interest. This appears to be an important step forward in our understanding of the formation and evolution of galaxies and at the same time demonstrates the power of the authors' sequential strategy for searching an enormous space for increasingly likely parameter values.

It may be helpful to illustrate my understanding of the authors' strategy in terms of standard statistical methodology using a simple problem. Suppose $Y_i \sim \mathrm{N}(\mu_i(\theta), \sigma_i^2)$ are independent for $i = 1, \ldots, n$, with each σ_i^2 known. The loglikelihood function is $L(\theta|Y) = -\sum_{i=1}^{n}(Y_i - \mu(\theta))^2/2\sigma_i^2$. If $\mu(\theta)$ is not overly complex, we can maximize L and consider its curvature or contours to make inference and learn about θ. We can also evaluate $L(\theta|Y)$ at its maximizer, $\hat{\theta}$, or values of θ near $\hat{\theta}$ to check whether the proposed Gaussian model is adequate for the data. If $L(\hat{\theta}|Y)$ is significantly smaller than we would expect we conclude that the model is inadequate. The authors consider a problem in which $\mu(\theta)$ is very complex, the likelihood can only be evaluated with substantial numerical effort, and standard optimization, quadrature, and sampling techniques are apparently impossible or impractical. Instead they search the parameter space by simply evaluating the objective function, $L(\theta|Y)$ in my simple example, at numerous values of θ. The evaluation points are then culled by thresholding on $L(\theta|Y)$. A new set of values of θ are selected in the newly discovered highest-likelihood region of the parameter space, the likelihood is reevaluated at these points, and the evaluation points are culled again using a more stringent threshold. This is repeated until a set of parameter values is obtained that adequately predict the observed data or until all possible value of θ have been eliminated by the likelihood threshold. Although the authors do not refer to the Gaussian loglikelihood function, the actual objective functions that they employ bare a remarkable resemblance to it. Using my notation, the *implausibility function* defined in (13) replaces the sum over i in $L(\theta|Y)$ by a maximization over i and the function in (16) would reduce to $L(\theta|Y)$ in the independent case. In both cases $\mu(\theta)$ involves emulation and the implausibility functions differ from the loglikelihood by a factor of $-1/2$. Thus, the authors aim to reduce implausibility as I aim to increase the likelihood. Having identified the set of parameter values that adequately predicts the

*Department of Statistics, University of California, Irvine, CA, mailto:dvd@ics.uci.edu

© 2010 International Society for Bayesian Analysis DOI:10.1214/10-BA524D

data, the authors graphically compared the corresponding predicted light curves with observed light curves and draw conclusions about the likely relationships among the parameters.

Employing the Common Statistical Framework for Computer Modeling. It is important to emphasize that the authors' strategy differs from standard statistical methods mainly from a computational point of view. Because the Galform computer model is highly complex, optimizing the objective or likelihood function (or sampling from the posterior) is apparently infeasible. As in a typical statistical problem, however, they aim to find the values of the parameter that result in the best fit to the data, being mindful that the best fit may not be good enough if the model is inadequate. These are the standard model fitting and model checking tasks that are a part of any sensible analysis, although they may be sliced and diced differently in different settings. Of course unique complications do arise with complex computer models. In many settings, for example, researchers aim to experiment with computer models. This involves first calibrating (i.e., fitting) the model by comparing its predictions with actual data and then using the calibrated model—often with the help of emulation and properly accounting for uncertainty in the fitted parameters—for prediction under alternate values of certain covariates. When collecting actual data under different experimental conditions is expensive or impossible researchers may use the calibrated computer model in place of actual experiments. Even this more complicated situation is analogous to prediction (or extrapolation!) in standard statistical terminology.

The computer modeling community's habit of employing different terminology tends to obfuscate the relationship between their techniques and standard tried-and-true statistical techniques. Rather than using ubiquitous terms like "likelihood function" or "discrepancy measure" they use "implausibility measure"; "model checking" and "model fitting" are referred to as "history matching" and "calibration"; and both "parameters" and "covariates" are simply called "inputs". This clash of terminology is certainly not unique to this paper. It appears in many computer modeling papers published in the statistical literature. Certainly there are many computer modelers who are not or are not primarily statisticians. But this can be said for Bayesians and other methodological groupings which do a better job of maintaining a unified basic framework for discussion. The lack of such a framework in computer modeling is a shame not only because it makes it much more difficult for the uninitiated statistician to learn about the sophisticated statistical techniques developed for these complex models, but also because it obscures the relationship between our rich library of statistical methods and their potential application to problems involving computer models. For example, the implausibility measure in (16) seems a much more natural choice than that in (13) when we realize that it is the objective function used to identify *likely* values of the parameter. Except for the sign, (16) is a *likelihood* function while minimizing the maximum descrepancy in (13) is employing a minimax criterion for parameter fitting. Of course, there may be computational reasons to prefer (13) - (15) when exploring the parameter space, but in the final analysis, on the face of things, (16) seems the better choice.

An even better case for bringing computer modeling terminology in line with that of statistics is that computer models are not always treated in isolation. A modern Bayesian statistician readily combines model components for different observed or latent quantities through hierarchical or multi-level models. These components may include parametric, non-parametric, multi-scale, and computer models, that are strung together into a unified model for a coherent statistical analysis. This job is made much easier if we focus on a common framework for working with these different models at least as much as we emphasize their subtle differences. And I emphasize that I am not saying that there are no differences, but rather that the commonalities are more important—certainly enough to justify use of common notation and vocabulary.

What Does it Mean to be Bayesian? In my first reading of the article, it was not clear to me why the word "Bayesian" appears in the title. Many of the standard hallmarks of a Bayesian analysis are absent. There is no specification of a prior distribution or computation of the posterior distribution. The Bayes Linear approach is based on Bayes Theorem, but by avoiding specification of the distributions involved, the authors miss what I believe to be one of the biggest benefits of a Bayesian analysis: a principled analysis that fully accounts for the complexities of the underlying distributions and avoids the old and often unrealistic Gaussian assumptions. Of course, this requires us to make certain assumptions about the distributions, but these assumptions are no more arbitrary than Gaussian assumptions and are clear for all the world to see and to evaluate for themselves.

The authors take a different approach. They take advantage of the ability of a Bayesian analysis to account for information from outside the data, such as the opinions of experts. Indeed much of their model is justified solely in terms of expert judgements. For example, they "judge" the experimental error to be uncorrelated with "true" physical system values. The model discrepancy, that is the difference between the "true physical system values" and the Galform model evaluated at the "actual" parameters, is judged to be uncorrelated with Galform evaluated at the "actual" parameters. Expert judgement is used to quantify the likely size of the model discrepancy along with its "rich covariance structure". This judgement is critical to the final analysis because it determines how large the residuals may be without bringing the overall model into question. Thus whether the parameter values from Wave 5 match the data sufficiently well to conclude that Galfrom is an adequate model of Galaxy Formation is determined by expert judgement on the model discrepancy.

Of course any statistical analysis requires subjective assessments. When non-Bayesians complain about our use of prior distributions, we rightly point to the assumptions involved in their specification of the likelihood function. In many ways it is a matter of art to weigh the subjective assumptions in a statistical model. Nonetheless, most Bayesians and non-Bayesians alike work hard to be cognizant of the assumptions inherent in their models, to employ careful model checking, and to investigate the sensitivity of their final analysis to their assumptions. In this case the authors use multiple runs of the Galform model with the help of the Millennium Simulation to access the effect of the fixed parameters and the necessary "specification of the arrangement of Dark Matter at all times in the development of the universe". The magnitude of further model discrepancy is judged with the help of an elicitation tool. While the authors are certainly going to great lengths to carefully quantify these subjective quantities, there is too much expert judgement for me to find the final results convincing. I worry that the expert assessments determine the final outcome. It seems that the magnitude of the model discrepancy alone is enough make or break the model checking and this is ultimately decided by an expert turning the knobs of an elicitation tool.

While I agree with the authors that "a Bayesian analysis has value largely because...it is an appropriate way to combine expert judgement and observations to give appropriate posterior judgements," in my view even a Bayesian analysis must work hard to minimize its assumptions and must be absolutely upfront about the impact of its subjective assessments on the final analysis.

If it Looks Like a Duck.... The authors justify their use of a Bayes Linear Analysis by noting that it allows them to avoid distributional assumptions and to base inference on the first two moments alone. They argue that it is much easier to specify these moments and that the posterior distribution may be "highly non-robust" to aspects of the distributions other than the first two moments and that these aspects cannot be "specified with confidence". They go on to note that "a full Gaussian specification for all of the relevant quantities would lead to similar updating formulae" and to evaluate the Bayesian linear analysis in terms of the best linear fit under squared error loss. The objective functions used to identify plausible values of the parameter given in (13)–(16) are all based on squared error loss and the Gaussian loglikelihood. The model forms given in (1) and (2) rely on additive uncorrelated error. Perhaps this is simple minded on my part, but I'm far less concerned with theoretical considerations such as the "infinite number of further joint orthogonality constraints as required by full probabilistic independence" over the authors' assumptions of no correlation than I am with what happens in the actual analysis. That is, how would the final analysis differ if a Gaussian distribution where specified for the full joint distribution? That there would likely be little or no difference is troubling because, as the authors note, the likelihood function under computer models of this sort typically exhibits an "extremely complex, multi-modal form". How then should we interpret the results of a seemingly Gaussian analysis?

The scientific objective. In my work with astronomers the ultimate goal is always to learn about the likely values of physical parameters of scientific interest. They are interested in the precise location, age, metal content, or distance of a particular source. Thus, I was quite puzzled by the authors' statement that

> Achieving an acceptable match, for a particular input choice x, does not mean that the model is "correct" or that a parameter choice which achieves the match corresponds to the "true" value of the parameters, but simply that this version of the model will have met the challenge of reproducing an important observational aspect of the galaxy formation study...

Of course all models are parsimonious summaries and are not meant to capture all of the complexity of the physical system. But we do hope that they capture enough of the important features of the system so that we have some capacity to meet our scientific objectives. In many statistical analyses the ultimate goal is prediction, and, in this case a black box that bears no resemblance to the actual generative system but nonetheless predicts well will meet the statistical objectives. I find this rarely if ever to be the case in astronomy. The goal is not to predict new universes, but to understand our own. If the authors really don't believe the model will shed light onto the actual physical world, what is the ultimate goal of this analysis?

The Final Analysis. In their careful analysis, the authors obtain a set of parameter values that adequately predict the observed light curves. These are used to learn about likely relationships among the parameters. The authors provide matrices of scatter plots using coloring to indicate the value of the objective function for each acceptable parameter value. (I believe they use the function in (16) that corresponds to a multivariate Gaussian loglikelihood.) A final step would be to weight each of the parameter values by its posterior density or perhaps likelihood and resample according to the weights. This would provide an approximate Monte Carlo sample form the posterior distribution. Alternatively, kernel density estimation could be used to construct a proposal distribution for use in an independence Metropolis-Hasting sampler. In either case, the resulting Monte Carlo sample should provide an adequate approximation to the posterior distribution if a Gaussian model is used in its formulation. On the other hand, if the more realistic model that the authors allude to as resulting in an "extremely complex multi-modal" likelihood is used, the approximation may not be as good. This may be detected in a highly skewed set of weights of a poorly mixing Markov chain and would indicate that the Gaussian-like assumptions of the article are critical. If the Gaussian-like assumption proves to be benign, the authors may be very closer to a fully Bayesian analysis.