Invited Discussion*

Steven N. MacEachern† and Juhee Lee‡

First, we would like to congratulate the authors for their development of a fast and efficient method of assessing sensitivity to the prior specification of the Dirichlet process mixture (DPM) and related models. Their techniques are widely applicable and will see much use. Their examples give us a taste of what can be done with their method in models that move beyond the DPM. We greatly enjoyed reading the paper.

The past 30-plus years have seen incredible growth in the use of Bayesian methods for data analysis. The initial growth was driven by the development of Markov chain Monte Carlo (MCMC) methods that allowed one to fit models of substantial complexity. This was quickly followed by a realization in the entire statistics community that Bayesian methods simply work better than classical methods in "high information" settings with clean data – settings where one can reliably write down a complete Bayesian model and can clearly specify the inference problem. In practice, this translates to settings where different analysts (or the same analyst on different days) would select similar sampling densities for the data and similar prior distributions for the parameters, including latent structures, to arrive at similar models, and would also choose similar loss functions to formalize the inference problem. It also relies on realism in the model, with choices based on an understanding of the phenomenon under study rather than computational convenience. In these settings, the mathematical theory that links optimal inference to Bayesian methods is borne out. Bayesian methods simply work better than methods that are distant from Bayes.

The success of Bayesian methods is not uniform. In "low information" settings, specification of the sampling density and form of the model are every bit as challenging for the Bayesian as for the classical statistician. For high or infinite dimensional models, specification of the prior distribution remains a challenge. Data of dubious quality have the potential to dramatically impact the final inference. MCMC methods are often slow and sometimes exhibit poor convergence. And analysts commonly adjust their model to make fitting it easier and quicker. These difficulties are not shortcomings of the analyst, but rather features of the low information-complex model setting. They set an agenda for research on Bayesian data analysis.

The centerpiece of this agenda is how to improve Bayesian data analysis. Here, we see three main threads. One focuses on diagnostics through the development of techniques to identify deficiencies in the model (whether sampling density, prior distribution or a combination of the two), to identify cases that do not accord with the model (as in outliers), and to identify cases that have a large impact on inference (influential cases).

*Supported by the NSF under grant numbers SES-1921523, DMS-2015428, and DMS-2015552.

†Department of Statistics, The Ohio State University, ORCID 0000-0003-4106-1232, snm@stat.osu.edu

‡Department of Statistics, University of California, Santa Cruz, ORCID 0000-0002-9787-3830, juheelee@soe.ucsc.edu

The second focuses on computational implementation. The third focuses on strategies to improve model specification and inference, with particular attention to robustness in the low information setting.

Giordano et al.'s delightful paper focuses on the first thread and is informed by the second. The authors' insight into the (in)effective use of Bayesian methods shines sharply through their paper. They consider a high/infinite dimensional setting where the prior distribution is specified through a rule-based strategy and where it would be difficult to place full confidence in any chosen rule. In this same setting, variational methods for fitting the model are far quicker than MCMC methods. Variational Bayes (VB) methods also generate the derivatives needed to examine the local sensitivity of features of the posterior distribution to changes in the prior distribution.

The developments in Giordano et al.'s paper parallel the development of local influence as a diagnostic method in classical statistics. Cook (1977)'s initial development of case influence examined the impact of individual cases on inference in the linear model. The main technique was case deletion. It led to Cook's distance, now a standard diagnostic summary in regression. Cook (1986) subsequently extended measures of influence to classical nonlinear models which, in the early-to-mid 1980's, were subject to the difficulties more recently experienced by MCMC methods. The models were slow to fit (via maximum likelihood) and one needed to be concerned with the numerical accuracy of the fits. With no clean analytical form, these difficulties rendered case deletion methods ineffective for the interactive data analysis that was being developed at the time. Instead, Cook turned to local influence, considering infinitesimal perturbations of case weights, and looked for big directional derivatives.

Cook's methods have been extended to the Bayesian setting, initially with pre-MCMC computation (e.g., Johnson and Geisser (1983)) and then with MCMC (e.g., Weiss (1996), Bradlow and Zaslavsky (1997), MacEachern and Peruggia (2000)). The approaches include both full case deletion and local influence (viz. Thomas et al. (2018)), and they cover a variety of inferences, from impact on the full posterior distribution to impact on marginal summaries of the posterior. While these methods are successful for linear models and low-dimensional nonlinear models the techniques are less effective for high-dimensional problems of the sort considered by Giordano et al.

Our first question for the authors is whether the techniques they develop can be adapted to assess local case influence. If so, is such an adaptation computationally feasible? The extension would provide the analyst with an additional tool to identify cases or sets of cases that have a large impact on inference.

Our second question concerns robust forms of the prior distribution. As described in the paper, DPM models are often used for clustering problems. Inferences on clustering, such as the number of clusters and co-clustering probabilities, are influenced by the prior specification. The parameter of the Dirichlet process (DP) may be split into two parts, the total mass parameter α and a base probability measure, $\mathcal{P}_{\text{base}}$. The distribution $\mathcal{P}_{\text{base}}$ generates cluster specific parameters, i.e., $\beta_k \overset{iid}{\sim} \mathcal{P}_{\text{base}}(\beta \mid \xi)$, where the β_k's are cluster specific parameters and ξ is the hyperparameter vector for $\mathcal{P}_{\text{base}}$. Jointly

with α, $\mathcal{P}_{\text{base}}$ influences inference on the clustering structure. For example, Bush et al. (2010) and Lee et al. (2014) studied the joint impact of α and the dispersion of $\mathcal{P}_{\text{base}}$ on posterior inference. In particular, for fixed α and a given dataset of size n, a DPM model tends to produce fewer clusters with larger sizes under more dispersed $\mathcal{P}_{\text{base}}$. In response to this phenomenon, Lee et al. (2014) developed a local-mass preserving prior distribution for Bayesian nonparametric (BNP) models that produces more stable inference for clustering. The central idea is to define "local mass" as the mass assigned by a measure to a region $\mathcal{L}$ of interest in the parameter space Ω_β prior to analysis and to jointly elicit $\mathcal{P}_{\text{base}}$ and α by holding the mass in $\mathcal{L}$ constant. In addition to providing more stable inference about clustering, this form of prior distribution stabilizes inference for other quantities such as estimation of cluster locations.

It would be of great interest to see whether the authors' method can be extended to perform a more comprehensive examination of model sensitivity, including sensitivity to the local mass $\mathcal{L}$. Does the authors' quick and automated tool to assess sensitivity of inference to the specification of α extend to prior distributions with a local mass structure? If so, does one form of prior distribution show greater robustness than the other?

Our final comment returns to data analysis. The standard Bayesian analysis requires a rigorous determination of three components; (i) the sampling distribution (likelihood function) for the observations, (ii) the prior distribution of the parameters and (iii) loss function for making inference (decision). Bayesian analysis strongly depends on the choice of these components, and it is essential to investigate the sensitivity of the procedure to perturbation of all three components. The loss function has traditionally received less attention than the prior distribution and sampling density, as it often has little impact in a low-dimensional parametric setting. However, the choice of the inference (loss) function is critical for BNP models due to their flexibility. For example, DPMs are good at accommodating local features of the data such as outliers. These cases may be captured as one or more clusters that depart from the general pattern of the data. Thus, an inference function that discounts the impact of the outliers on the overall analysis can be more desirable than traditional inference functions (e.g., the quadratic loss function and the 0-1 loss function) for robust decision making (Lee and MacEachern, 2014). The inference function does not "wash out" as the sample size grows. This is similar to the parameter α not washing out for clustering in DPM models. We see scope for the development of inference functions that target a sensible summary of the posterior (or predictive) distribution and that lead to stable inference as the prior and sampling density are varied. Do the authors have a sense of whether a summary such as "number of clusters exceeding a given size" tends to be more robust than the simple "number of clusters"? Do the authors know of systematic ways to create more robust summaries of clusters?

In keeping with the level innovation in this work, the paper opens a host of questions. Those above seem, to us, to walk the tightrope of computational feasibility that leads from the questions we can answer with our written model to those we would answer with infinite resources. We close our discussion here, congratulating the authors on an interesting paper that develops a technique that will undoubtedly see heavy use.

References

Bradlow, E. T. and Zaslavsky, A. M. (1997). Case influence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics*, 6(3):314–331. 322

Bush, C. A., Lee, J., and MacEachern, S. N. (2010). Minimally informative prior distributions for non-parametric Bayesian analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):253–268. MR2830767. doi: https://doi.org/10.1111/j.1467-9868.2009.00735.x. 323

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19(1):15–18. MR0436478. doi: https://doi.org/10.2307/1268249. 322

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 48(2):133–169. MR0867994. 322

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, 78(381):137–144. MR0696858. doi: https://doi.org/10.1080/01621459.1983.10477942. 322

Lee, J. and MacEachern, S. N. (2014). Inference functions in high dimensional Bayesian inference. *Statistics and Its Interface*, 7(4):477–486. MR3302376. doi: https://doi.org/10.4310/SII.2014.v7.n4.a5. 323

Lee, J., MacEachern, S. N., Lu, Y., and Mills, G. B. (2014). Local-mass preserving prior distributions for nonparametric Bayesian models. *Bayesian Analysis*, 9(2):307–330. MR3216998. doi: https://doi.org/10.1214/13-BA857. 323

MacEachern, S. N. and Peruggia, M. (2000). Importance link function estimation for Markov chain Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 9(1):99–121. MR1819867. doi: https://doi.org/10.2307/1390615. 322

Thomas, Z. M., MacEachern, S. N., and Peruggia, M. (2018). Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models. *Journal of the American Statistical Association*, 113(524):1669–1683. MR3902237. doi: https://doi.org/10.1080/01621459.2017.1360777. 322

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(4):739–750. MR1410188. 322