Bayesian Analysis (2013) **8**, Number 2, pp. 303–310

Comment on Article by Müller and Mitra

Bradley P. Carlin[*] and Thomas A. Murray[†]

Congratulations to the authors on a fine and comprehensive review of nonparametric Bayesian (BNP, or NPB?) inference! The authors offer a whirlwind review of an enormous field that has seen explosive growth since the advent of MCMC methods for Bayesian computing; indeed, the two fields rose to prominence simultaneously around 1990, and enjoyed substantial cross-pollination (see Escobar and West, 1998, for a review). One of us (BPC) admits to being somewhat surprised to be asked to serve as a discussant for this paper, since if he has any professional reputation at all, it is that of an unapologetic Bayesian parametrician who prefers `BUGS` implementions, with theoretical properties simulated via repeated calls from `R` using `BRugs` or `rjags`. But the editors insisted this is precisely what they were looking for, and after reading the paper we agree with them: this is a very nice paper that is in need of a crabby parametrician or two to fuss about its breezy claims of heretofore-unimagined modeling freedom and insight, especially when there is often relatively modest payoff for the effort expended. As such, our outlook here is like an asymptotician's wary view of the Gibbs sampler in 1990, or any frequentist's view of the phrase "applied Bayesian statistics" before 1980. While the progressive Gibbs-applied viewpoint prevailed in both those cases, the other ideas hardly vanished; asymptotic approximations are even cool again thanks to the emergence of the `INLA` package (c.f. Rue et al., 2009). Thus we hope the reader will grant us our slightly peevish, "you-kids-get-off-my-lawn" tone so that we may fulfill our Prof. Cranky Pants role, as the editors (and we) believe we must.

BNP is of course all about flexibility; in their very first paragraph, the authors speak glowingly of "allowing for a richer and larger class of models." Indeed this has been a central theme of all Bayesian modeling since 1990, so publishing this in *Bayesian Analysis* is preaching to the choir to be sure (and what a choir; the authors' reference list is long and only scratches the BNP surface). But the ways in which this flexibility manifests are sometimes mysterious, and other times easily mimicked by a carefully considered yet much simpler parametric model. Later in their introduction, the authors worry that, "Restriction to a parametric family can mislead investigators into an inappropriate illusion of posterior certainty." In our admittedly limited experience, a much more common problem in Bayesian modeling is accidental *over*-parametrization of a previously-understood model, resulting in poor identifiability and associated slow MCMC convergence. Indeed, the recent explosion in BNP research has led to the publication of a few BNP models before their utility had been established for even a single real dataset. In our opinion, mere flexibility for flexibility's sake is not enough; the flexibility must be both well-understood and routinely controlled. In this regard, we were puzzled by the authors' Figure 2(a), which overlays their BNP results with arguably the world's most popular nonparametric estimator, the Kaplan-Meier empirical survival curve. The former are a poor match to the latter for $Y > 2.5$, but this discrepancy is

[*] Division of Biostatistics, University of Minnesota, Minneapolis, MN, brad@biostat.umn.edu
[†] Division of Biostatistics, University of Minnesota, Minneapolis, MN, murra484@umn.edu

DOI:10.1214/13-BA811A

not mentioned. In fact, there seem to be some discrepancies with the labeling of the data posted to the authors' website (the censoring indicator and the study arms appear to be mislabeled), so we were unable to analyze these data ourselves. Nevertheless, based on the Kaplan-Meier survival curves shown in the paper, there do not appear to be any observed events in either treatment arm more than 7.5 years after treatment. With only censored information near year 8, it is difficult to say whether the "bump" in the p.d.f. to the right of $T = 8$ in Figure 2(b) is truly there, or merely an artifact of the cure-rate BNP model employed.

This brings us to our second point, which is the inherent difficulty in explaining the results of some BNP analyses to ourselves, our students, and our substantive area colleagues. While parameters are unobservable creatures, their presence in models does often help our intuition about what is going on scientifically, and these colleagues often prefer such models and their relatively straightforward interpretations. But with BNP modeling, much of this comfort level is lost, and replaced with piles of mathematical notation that is largely impenetrable to non-statisticians. Indeed, the job the authors have undertaken here (review all of BNP methods in a single paper) is inherently impossible; there is just too much out there. This has led some sections of the paper to look rather "bare bones"; we suspect even the most savvy and diligent but BNP-naive statistics PhD students would have a hard time grasping Polya trees and their mixtures by reading only Section 2.2. Of course, the many researchers working in this area know this, and have sought to develop `R` packages like `DPpackage` to help out. But it would still be easy to worry that our beloved Bayesian inference engine had now moved irrevocably to "black box" status if we were to buy into a purely BNP approach. Some recent introductory, "Bayes-and-`BUGS`" books (Kruschke, 2011; Lunn et al., 2012) go part of the way in this direction, introducing Bayes rule mostly via example, and encouraging students to rely as much on the software and their intuition about Bayesian learning as on mathematical equations or calculus. But at least these models are still accessible to most students and scientists with quantitative but not strongly mathematical backgrounds.

Even if we grant that the flexibility of BNP is well worth considering and its complexity likely to be worth the payoff, there is still the matter of comparing such results to those of parametric methods likely to be competitive (according to some pre-agreed performance metric). In addition to Figure 2, another setting where the authors did not do this is their very first example, the analysis of the zero-censored discrete data in Table 1. Here the authors reject a finite mixture of Poissons as likely to produce "misleadingly precise inference" (repeating their earlier worry), and laud the BNP method's ability to deliver a model-based yet appropriately imprecise estimate of $p(F(0)|\mathbf{y})$. While it is true that the 95% BNP credible interval for $F(0) = P(Y_i = 0)$ is indeed wider (suggesting values from around 0.3 to 0.7) than any simple parametric alternative would produce, we were struck by the *narrowness* of the BNP interval for $F(1) = P(Y_i = 1)$ (which is consistently estimated to be near 0.3), and its poor agreement with the empirical proportion of 1's among the nonzero values ($37/55 = 0.67$) apparent in Figure 1(a). Obviously the BNP model believes that a lot of zero observations were missed, and the extent to which this is actually true determines the model's performance. Of course

	Hurdle Poisson		ZIP	
q prior:	Beta(70,30)	Beta(1,1)	Beta(70,30)	Beta(1,1)
q	0.81 (0.74, 0.86)	0.98 (0.94, 1.0)	0.81 (0.74, 0.87)	0.98 (0.94, 1.0)
λ	0.86 (0.58, 1.18)	0.86 (0.59, 1.18)	1.49 (1.19, 1.83)	1.49 (1.19, 1.83)
$F(0)$	0.19 (0.14, 0.26)	0.02 (0.00, 0.06)	0.38 (0.30, 0.46)	0.24 (0.17, 0.33)
$F(1)$	0.51 (0.41, 0.61)	0.62 (0.51, 0.73)	0.27 (0.23, 0.30)	0.33 (0.29, 0.36)
$\overline{D}$	129.5	107.7	169.1	147.4
p_D	1.07	1.02	1.08	1.02
DIC	130.6	108.7	170.2	148.4

Table 1: Posterior summaries and model choice statistics, hurdle Poisson and ZIP models, for two different beta priors on q.

we don't know the true $F(0)$ and it is largely inestimable by design, but we began to wonder how much one could learn from a comparison of standard parametric models in this setting. As such, we used `OpenBUGS` to fit two simple alternatives, a *hurdle Poisson* model (see e.g. Neelon et al., 2010),

$$F(k) = P(Y_i = k|\lambda, q) = \begin{cases} 1-q\,, & k=0 \\ q\frac{e^{-\lambda}\lambda^k}{k!(1-e^{-\lambda})}\,, & k=1,2,\ldots \end{cases} \tag{0.1}$$

and a *zero-inflated Poisson (ZIP)* model (Lambert, 1992),

$$F(k) = P(Y_i = k|\lambda, q) = \begin{cases} (1-q)+qe^{-\lambda}\,, & k=0 \\ q\frac{e^{-\lambda}\lambda^k}{k!}\,, & k=1,2,\ldots \end{cases} \tag{0.2}$$

where in both cases $\lambda > 0$ and $q \in (0,1)$. Note that model (0.1) mixes a point mass at 0 with a truncated Poisson for the nonzero observations, while model (0.2) mixes the point mass with an *untruncated* Poisson. These are both relatively simple, 2-parameter models, yet they lead to results that are surprisingly different from each other and from those in Figure 1.

Our implementations used the `dcat` function in `OpenBUGS`, and ran for 30,000 iterations after a 1000-iteration burn-in period. Table 1 provides the DIC-based model choice and posterior parametric summaries from models (0.1) and (0.2), where we use a vague prior for λ and compare the effects of two priors for the point mass parameter q: an informative $Beta(70,30)$ (that expects a point mass at 0 of somewhere between 0.2 and 0.4), and a noninformative $Beta(1,1) = Unif(0,1)$ distribution. We see that the prior on q is quite influential (not surprising in this missing data setting), but its influence is little affected by the choice of likelihood. Conversely, the Poisson parameter λ is quite different between the two likelihoods, but unaffected by the choice of q prior. Yet when we turn to the fitted probabilities of 0 and 1, $F(0)$ and $F(1)$, we see very large differences across both models and priors. The hurdle Poisson consistently leads to higher $F(1)$ values, and correspondingly lower $F(0)$ values, than the ZIP model; we note that the 95% Bayesian credible intervals for these quantities do not even overlap across models and priors. The DIC scores suggest the hurdle model's greater fidelity to the observed proportion of 1's in the dataset (as compared to both ZIP and DPM) leads to better fit statistics. Perhaps equally interesting, the total range for $F(1)$ suggested across all four parametric models extends from 0.23 to 0.73, a much, much wider range than that indicated by the DPM model. For us, this calls into question the authors' preference for the DPM model here: its fit for $F(1)$ is poor, and any alleged "flexibility compared to a simpler parametric family" does not manifest for this model quantity; mixing across the 4 simple parametric models in Table 1 would produce an even wider credible interval.

The authors *do* consider parametric alternatives in Figure 4, which compares BNP and parametric results, and finds the latter wanting due to its strong respect for the "good vs. intermediate prognosis" classification system apparently parametrically imposed on this model. But here again, we do not feel that the authors have fairly illustrated the flexibility parametric approaches can provide. Here, estimation of the probability of response π_i across a variety of sarcoma subtypes is of primary interest, but it is somewhat unclear *a priori* how similar each subtype is to the others aside from the prognosis. We investigated three parametric logistic models that facilitate various amounts of between-subtype borrowing of strength. Each of these models uses the paper's binomial likelihood, and link function

$$\text{logit}(\pi_i) = \alpha + \gamma x_i + \theta_i \ ,$$

where $x_i = \{-1, 1\}$ for good and intermediate prognosis, respectively. For all three parametric approaches, we model the random effects as $\theta_i|\tau \overset{iid}{\sim} N(0, \tau)$, and assign vague priors to α and τ. We then form three model variants simply by altering the prior on γ. First, we fit an *exchangeable* model that borrows across all subgroups simply by taking $p(\gamma) = 0$, i.e., by not permitting a prognosis effect at all. Second (and at the other extreme), we form a *nonexchangeable* model by placing a vague prior on γ having mean 0 and near-zero precision, a construction that allows the data to dictate the effect of prognosis on the mean of the π_i's. Finally, we also consider a *partially exchangeable* model that shrinks the estimate of γ toward zero by modeling $\gamma|\tau_0 \sim N(0, \tau_0)$ and placing a vague gamma prior on τ_0, calibrated to deliver a prior 95% credible interval of $(0.0033, 300)$ for the odds ratio for good prognosis versus intermediate prognosis.

For all three models, we again used *OpenBUGS* to obtain 10,000 posterior draws from 2 parallel chains, each following a generous 45,000-iteration burn-in. The resulting posterior estimates were robust to modest changes in our prior distributions. Our Figure 1 displays the 95% posterior credible intervals for the three parametric models. Note that our very simple exchangeable model (solid lines) produces interval estimates virtually indistinguishable from those from the BNP model (solid lines in the authors' Figure 4), with the exception of the upper limit for Rhabdo ($i = 2$). However, both of these models appear to preclude very low event rates for the two good prognosis subtypes, despite strong evidence to the contrary in the data ($y_i = 0$ for $i = 1, 2$). By contrast, our partially exchangeable model (dotted lines in our figure) fixes this problem, and nicely illustrates that parametric models can in fact deliver sensible results that are intermediate to the two extreme cases.

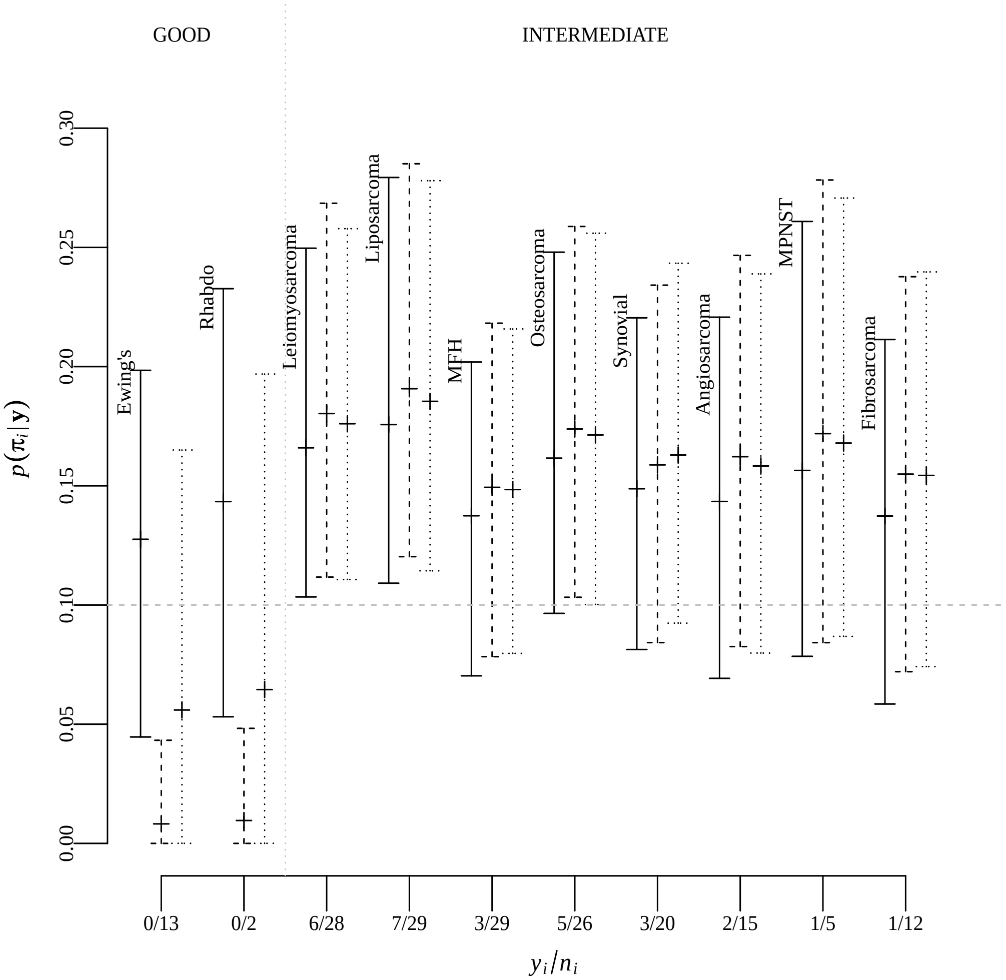

Figure 1: Central 90% posterior credible intervals of success probabilities π_k for each sarcoma subtype under three types of parametric logistic models: exchangeable (solid), nonexchangeable (dashed), and partially exchangeable (dotted). The central marks ("+") denote the posterior means. The partially exchangeable model produces results intermediate to those of the other models, which make diametrically opposed *a priori* assumptions regarding exchangeability of the two prognosis groups.

Of course, parametricians use fairly complex, partially exchangeable random effects structures all the time; our Minnesota colleague Jim Hodges calls them "smoothers" (so a CAR prior is a "spatial smoother", etc.). DPM priors offer a particular, very flexible kind of smoother which may well be attractive in various settings. But parametric smoothers may be able to do a similar job, in a way that is more accessible for practitioners, and without an infinite number of parameters. For instance, recent research in what are called commensurate prior models (Hobbs et al., 2011, 2012) offer a parametric compromise to the stark choice between "separate analysis" and "pooling" the authors mention (and just illustrated in our Figure 1). Recently, commensurate priors have been used in clinical trial survival modeling, both for drug trial design (Hobbs et al., 2013) and post-market medical device surveillance (Murray et al., 2013).

Well perhaps this is enough for one cranky discussion! There is certainly much to like in this paper; the effectiveness and broad utility of DPM models is hard to deny, and we are very fond of the authors' writeup of nonparametric regression in Section 4, which nicely clarifies that BNP technology may apply to the regression mean structure, the underlying error distribution, or (provided care is taken) both. We also liked the mixed approach taken in Section 5, which represents a modern view of modeling wherein some key elements (say, the non-linear mixed-effects model in Example 7) are specified in traditional parametric terms, but other elements seek full nonparametric flexibility. To belabor our earlier MCMC analogy a bit further, all in all the current situation reminds us of the slow resolution of the Bayes-frequentist controversy over the past decade, in which Bayesians moved more toward default priors and away from hardcore subjectivism, while frequentists added basic MCMC-Bayes tools to their kitbags in order to take advantage of the newfound modeling freedoms their Bayesian brethren kept buzzing about. Indeed, at the 2005 Joint Statistical Meetings right here in Minneapolis, then-ASA President Brad Efron (who is actually from St. Paul) essentially argued that the disagreement was already over; the two sides had come together, both recognizing the advantages in the other's argument, and borrowing the other side's best stuff as appropriate. Perhaps 2015 will see the merging of parametric and nonparametric Bayesian inference in the same way, with even cranky guys like us routinely using both approaches. OK, maybe 2025.

Acknowledgment

The work of both authors was supported in part by NCI grant R01-CA157458.

Additional References

Escobar, M.D. and West, M. (1998). Computing Bayesian nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, *Lecture Notes in Statistics*, **133**, 1–22.

Hobbs, B.P., Carlin, B.P., Mandrekar, S., and Sargent, D.J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, **67**, 1047–1056.

Hobbs, B.P., Carlin, B.P., and Sargent, D.J. (2013). Adaptive adjustment of the randomization ratio using historical control data. To appear *Clinical Trials*.

Hobbs, B.P., Sargent, D.J., and Carlin, B.P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, **7**, 639–674.

Kruschke, J. (2011). *Doing Bayesian Data Analysis.* New York: Academic Press, 2011.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D.J. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis.* Boca Raton, FL: Chapman and Hall/CRC Press.

Murray, T.A., Hobbs, B.P., Lystig, T.C., and Carlin, B.P. (2013). Composite Kaplan-Meier and semiparametric commensurate Bayesian models for post-market medical device surveillance with historical survival information. Research Report 2013–004, Division of Biostatistics, University of Minnesota.

Neelon, B.H., O'Malley, A.J., and Normand, S.-L.T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, **10**, 421–439.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.