Bayesian Analysis (2013) **8**, Number 2, pp. 323–356

Contributed Discussion on Article by Müller and Mitra

Comment by Murray Aitkin[1] and Julia Polak[2]

Space restrictions limit our discussion to the first example.

The example analysis extrapolates from a zero-truncated observed count sample to predict the zero count. Prediction outside the data range is always hazardous. As the authors note, a general multinomial distribution on the observed data cannot predict the zero count: a parametric model is essential for this, with the consequent *strong model-dependence* of the prediction.

Nakatani and Sato (2008) have given a survey and discussion of the zero-truncated Poisson, negative binomial and other discrete distributions for this extrapolation, and a Bayesian analysis of the Poisson and negative binomial distributions can be found in Vergne, Calavas, Cazeau, Durand, Dufour and Grosbois (2012). These authors point out the limitation of the small sample size of their collected data that prevents them from fitting more complicated models. Moreover, they explain why using alternative nonparametric estimates is not suitable.

They used conventional parametric Bayesian methods, not the Dirichlet Process (DP). We follow their analysis for the first example. With only four counts, the data could be analysed by a truncated Poisson(μ) distribution. The MLE of μ is 0.86, and with a flat prior on μ the posterior distribution of μ is easily computed. The median is 0.876 and the 95% central credible interval is (0.60, 1.22). Transforming from μ to $55/(1-e^{-\mu})$ gives a posterior (Figure 1) for the total number N of T-cell types very close to the authors' Figure 1(b), with median 94 and 95% central credible interval (78,122). What additional information does the DP analysis provide? The authors aim to find a modeling approach between a "misleadingly precise" parametric model like this one, and a fully general multinomial model which could not provide information about the unobserved zero class.

The Dirichlet process with a Poisson base mass function leads to a truncated mixed Poisson distribution for the observed data. With only four support points in the data, no more than two components can be identified from the mixture likelihood, with two extra parameters over the truncated Poisson. The authors give no details of the complexity of their DP model, so it is unclear how it is related to the truncated Poisson or two-component mixed Poisson models, or to the truncated negative binomial distribu-

[1]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia
[2]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology (MEGA), University of Melbourne, Melbourne, Australia

 DOI:10.1214/13-BA811D

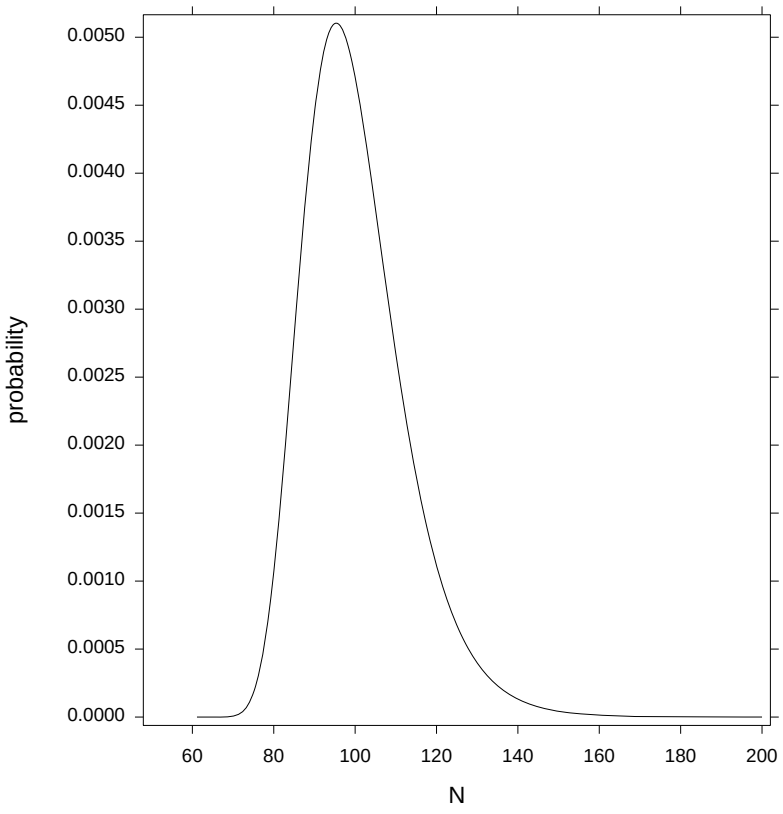

Figure 1: Posterior for number of T-cell types

tion which has one extra parameter. It seems therefore that the DP mixture leads to something very close to the single truncated Poisson distribution.

Is a more complex model needed? The truncated Poisson can be compared with the saturated multinomial using the posterior deviance distributions of the competing models (Aitkin 2010). The comparison of deviance cdfs for the truncated Poisson and multinomial shows that they cross around the 70th percentile – there is no strong preference for the more complex multinomial over the truncated Poisson. This supports the adequacy of the truncated Poisson, when compared to the multinomial.

We see an analogy between the DP and the variance component model for small area estimation. The variance component model expresses the variation in outcome y for some given kernel density $f(y \mid \theta)$ across k small areas j, through a random parameter a_j whose variance σ_A^2 determines the extent of the variation. The DP expresses the variation in outcome y for some given kernel density $g(y \mid \theta)$ through a "spawning" parameter α whose value determines the unobserved number k of components in a k-component mixture of the g distributions modeling y.

The important difference in these models is that the k small areas are known, while the k components are not. This means that the variance component σ_A^2 is estimable through the model likelihood, while the spawning parameter α is not: it has to have an informative prior to provide an analysis. The number of components k *is* however estimable through the likelihood.

We therefore see little point in the additional complexity of the DP analysis *for single samples*, in which the spawning parameter is unidentifiable: the mixture of kernel densities can be analysed by other Bayesian methods. Aitkin (2010) gives an extended discussion of the normal mixture model for the well-known galaxy data set.

Whether or not embedded in the DP, the authors' analysis does not in any way support their zero class proportion inference – many other distributions (like the truncated negative binomial) could be adequate as well, and could give a different zero class probability inference.

References

Aitkin, M. (2010) *Statistical Inference: an Integrated Bayesian/Likelihood Approach*. Boca Raton, Chapman and Hall/CRC Press.

Nakatani, T. and Sato, K. (2010) Truncation and endogenous stratification in various count data models for recreation demand analysis. *Journal of Development and Agricultural Economics* 2, 293-303.

Vergne, T., Calavas, D., Cazeau, G., Durand, B., Dufour, B. and Grosbois, V. (2012) A Bayesian zero-truncated approach for analysing capture-recapture count data from classical scrapie surveillance in France. *Preventive Veterinary Medicine* 105, 127-135.

Comment by Julyan Arbel[3] and Bernardo Nipoti[4]

In this discussion we focus on density estimation and show that BNP models naturally provide a tool that is fairly stable under rescaling of the data. Müller and Mitra deal with the flexibility of BNP models and show, through some examples, that their use can be advantageous in common inference problems. As for density estimation, the paper describes the DPM model by means of an application to inference on T-cell diversity, where the observations are counts. The specific nature of the dataset ensures that the scale of the data is not an issue. Nonetheless this is a ubiquitous concern in density estimation problems with observations from continuous distributions. Clearly, it is desirable that the estimates are not significantly affected by a rescaling of the data. A closely related problem refers to the estimation of multidimensional densities in spaces where different axes represent quantities with different physical dimensions. There is not a natural way to define a metric on the product space and scaling constants need to be set in order to relate units along different axes. This scenario arises, for example, with astronomical observations consisting of position and velocity of stars (e.g., Ascasibar and Binney 2005). Although we are not aware of existing BNP literature where this problem is directly investigated, it is worth mentioning that, as a matter of fact, BNP models have been used for density estimation in non–commensurable spaces. For example, both Müller et al. (1996) and Hanson (2006) analyse the well-known ozone dataset and, by means of DPM and MPT models respectively, deal with the problem of estimating multivariate densities in, e.g., radiation and ozone concentration product space. In the next section we illustrate, through a simulation study, that the flexibility of the DPM model provides a natural answer to the problem of estimating densities in non–commensurable spaces.

We investigate the performance of location-scale DPM models with multivariate normal kernels (introduced in Müller et al. 1996) for density estimation through the following synthetic example. We generate bivariate samples $\mathbf{D}^{(n)} = (\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})$, of size $n \in \{50, 100, 150, 200\}$, from the mixture of two normals:

$$\frac{1}{3}\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right) + \frac{2}{3}\mathcal{N}\left(\begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}0.7 & 0\\0 & 0.7\end{bmatrix}\right).$$

The true density f and a scatter plot of 100 observations are shown in Figure 5. Then we consider rescaled data $\mathbf{D}_c^{(n)} = (\mathbf{X}^{(n)}, c\mathbf{Y}^{(n)})$ with varying scale parameter c. We use a DPM model to estimate f, conditional on each sample $\mathbf{D}_c^{(n)}$, and we let $\hat{f}_c^{(n)}$ denote the estimated predictive distribution. Simulations are done by using the R package `DPpackage` (see Jara et al. 2011) (10,000 iterations with a 5,000 burn–in period); the prior specification we have set is standard and, importantly, does not take into account the scale of the data. As a first argument in support of the stability of the model with respect to rescaling, we show in Figure 5 the estimates obtained for $n = 100$ and two scales, $c = 0.1$ and $c = 10$.

[3]CREST, Université Paris-Dauphine, France
[4]University of Turin, Italy

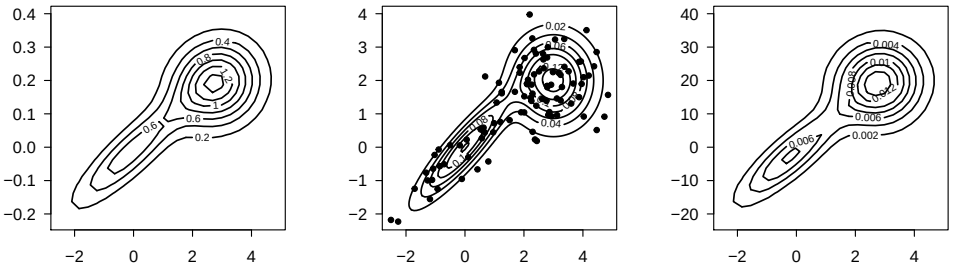

Figure 2: *(Middle)* Contour of the true density f and scatter plot of 100 observations. *(Left and right)* Contour of the estimates $\hat{f}_c^{(100)}$ for $c = 0.1$ and $c = 10$ respectively.

Additionally, for each n and $c = 10^{3k}$, where $k \in \{-2, \ldots, 2\}$, we summarize in Table 5 the fit of the estimate by computing the integrated squared error (ISE) for $\hat{f}_c^{(n)}$ suitably rescaled, that is

$$\mathrm{ISE}(\mathbf{D}_c^{(n)}) := \int_{\mathbb{R}^2} \left(c\hat{f}_c^{(n)}\left(x, y/c\right) - f(x,y) \right)^2 \mathrm{d}x\mathrm{d}y.$$

It is apparent that the fit of $\hat{f}_c^{(n)}$ is not heavily affected by the choice of c. This feature is even more evident when the sample size is large. It is worth stressing that the estimates we got are pretty stable even when the model is tested on data severely rescaled (e.g. $c = 10^{-6}$ and $c = 10^6$).

$n \backslash c$	10^{-6}	10^{-3}	1	10^3	10^6
50	4.73	4.77	4.87	5.25	5.24
100	2.29	2.27	2.25	2.68	2.65
150	1.90	1.92	1.93	2.17	2.35
200	1.07	1.07	1.06	1.13	1.17

Table 1: $10^3 \times \mathrm{ISE}(\mathbf{D}_c^{(n)})$ for varying data size n *(in rows)* and scale c *(in columns)*.

This toy example suggests that the flexibility of DPM models makes them good candidates for dealing with a whole range of density estimation problems for which there is not a univocal scaling of the data.

References

Ascasibar, Y. and Binney, J. (2005). "Numerical estimation of densities." *Monthly Notices of the Royal Astronomical Society*, 356(3): 872–882.

Hanson, T. (2006). "Inference for mixtures of finite Polya tree models." *Journal of the American Statistical Association*, 101(476).

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). "DPpackage: Bayesian Semi- and Non-parametric modelling in R." *Journal of statistical software*, 40(5): 1.

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83(1): 67–79.

Comment by Bertrand S. Clarke[5] and Gregory E. Holt[5]

We argue that the authors' focus on nonparametric Bayes estimation, despite being well executed, has led them to neglect the topic of nonparametric Bayes testing – a topic many non-statisticians think is just as important as estimation. Leaving aside whether estimation or testing is more important, our point here is that the arguments in favor of NPB from a testing perspective appear to have been neglected in general. As noted by Tokdar et al. (2010) 'The Bayesian literature on these testing problems is still rather meagre, unlike the case of nonparametric estimation...' Despite Borgwardt and Ghahramani (2009) and Holmes et al. (2012) our literature search did not turn up much evidence to invalidate this observation. So, let us give a class of settings where NPB hypothesis testing is likely to be better than parametric Bayes testing or Frequentist testing. We will focus on testing the equality of two distributions.

Consider the following thought experiment. A scientist is interested in conducting a clinical trial enrolling patients with end stage cancer who are otherwise out of treatment options. Despite the need for comparisons to placebo based control groups, clinical trialists realize patients do not enroll in studies where they may receive a placebo and therefore most of these trials remain uncontrolled. Researchers often rely on historical controls despite their known deficiencies.

As an alternative, to study therapeutic modalities in patients with terminal diseases, researchers could enroll patients only seen in clinic on one defined day while creating a control group formed from patients satisfying the same inclusion/exclusion criteria but seen on an alternative clinic day. We refer to this sort of control group as 'virtual' since it is constructed artificially after the treatment group is enrolled. The dependence between the treatment group and the virtual control group only comes from the inclusion/exclusion criteria and from matching the distribution of the baseline variables (described below). Such virtual control groups should exhibit the same outcome variable, here overall survival denoted Y, and Y should be a function of the baseline variables for both the treatment and virtual control groups. In this procedure, the virtual control group corresponds to patients receiving standard of care therapy so any differences between treatment and control would suggest a treatment effect.

Although placebo controlled randomized trials would still be preferable, in settings involving patients who typically avoid placebo controlled trials, this clinical trial design may permit better comparisons than historical controls that do not take into account current treatment practices or characteristics of the local population and treating physicians. In these contexts, NPB testing of the equality of the distribution of the baseline variables would be a better way to verify that a candidate virtual control group will provide a suitable comparison for a treatment group than Frequentist or parametric Bayes testing would be. At root, this follows because Bayes testing is better than Frequentist testing, see Berger and Bayarri (2004), Berger (2003), and M. Eaton (2013) among others, and nonparametric testing is more flexible than parametric testing.

To set up this testing problem, let us assume that all patients seen by a physician

[5]Department of Medicine, University of Miami, Miami, FL

on a day of experimental enrollment (say Tuesday) or on a day of virtual control group formation (say Thursday) have had the same basline tests. Now, in principle, we can compare the baselines of the patients in the Tuesday group with a collection of Thursday patients that we can use to form a virtual control group. More formally, suppose the baseline measurements for the treatment group are represented as $\mathbf{X} = (X_1, \ldots, X_K)^T$ and we have n outcomes $\mathcal{D}_T = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. To form a 'virtual control group' let $\mathbf{X}'$ be the same variables as $\mathbf{X}$ but measured on the Thursday patients and let $\mathcal{D}_C = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}$ be the resulting set of baseline measurements. The question is how to choose $\mathcal{D}_C$ so that we can compare the corresponding $Y_1, \ldots, Y_n$ from the treatment group with the $Y'_1, \ldots, Y'_n$ from the control group.

One way to formulate this is as a hypothesis test. Let P be the distribution of $\mathbf{X}$ and let Q be the distribution of $\mathbf{X}$. We want to test

$$\mathcal{H}_0 : P \neq Q \quad \text{vs.} \quad \mathcal{H}_1 : P = Q. \tag{1}$$

It is natural to use the Bayesian formulation. Foundationally, Bayesian techniques are not probabilistic in the data on which one conditions, see Chen (1985) Sec. 3.1. Specifically, the conditioning data need only form a well-defined deterministic sequence. So, it is legitimate to search the Thursday patients to find the ones that will give a $\mathcal{D}_C$ that lets us reject $\mathcal{H}_0$, i.e., mimics $\mathcal{D}_T$ well enough that the posterior probability of the null is small enough.

The NPB solution is clear: Find a nonparametric prior distribution for the pair (P, Q), for instance a bivariate DP as described in Walker and Muliere (2003) or a bivariate MDP as in the present paper. Now, reinterpeting (1) as

$$\mathcal{H}^*_0 : d(P, Q) \geq \epsilon \quad \text{vs.} \quad \mathcal{H}^*_1 : d(P, Q) < \epsilon, \tag{2}$$

for some distance d and writing the prior as W, the Bayes test is based on

$$\frac{W(d(P, Q) \leq \epsilon | \mathcal{D})}{W(d(P, Q) > \epsilon | \mathcal{D})} \tag{3}$$

where $\mathcal{D} = \mathcal{D}_T \cup \mathcal{D}_C$. If (3) is large enough then we are led to accept the alternative in (2) and therefore use $\mathcal{D}_C$ as a 'virtual control group' for inference on Y and Y'.

What would the nonparametric Frequentist solution be? First, (2) would be harder to test than (1), so let us focus on (1). Frequentist Neyman-Pearson testing treats the hypotheses asymmetrically and familiar two sample forms of tests such as Kolmogorov-Smirnov, the Anderson-Darling test, and the Cramer-von Mises test treat $\mathcal{H}_1$ vs. $\mathcal{H}_0$, the reverse of (1). To adapt such a test statistic to our present case requires that the null be decomposed into a series of nulls that can be tested separately and then put together by some kind of multiple comparisons procedure. That is, write

$$\{P \neq Q\} = \cup_{j=1}^{J} B((P_j, Q_j), \eta) \cup S \tag{4}$$

where $B((P_j, Q_j), \eta)$ is a collection of balls of radius $\eta > 0$ and $S = [\cup_{j=1}^{J} B((P_j, Q_j), \eta)]^c$ is a set of pairs of distributions deemed to be so far from the 'line' of distributions $P = Q$

that they can be ignored. Now it is enough to consider the J composite vs. composite tests $\mathcal{H}_{0,j} : (P, Q) \in B((P_j, Q_j), \eta)$ vs. $\mathcal{H}_1 : P = Q$. However, if η is small enough then

$$\mathcal{H}_{0,j} : (P, Q) \in B((P_j, Q_j), \eta) \approx \mathcal{H}^*_{0,j} : (P, Q) = (P_j, Q_j),$$

and for each j we can reduce $\mathcal{H}_1$ to $\mathcal{H}_{1,j} : (\tilde{P}_j, \tilde{Q}_j) = \arg\min_{P=Q} d((P_j, Q_j), (P, Q))$. So, to test (1), it is approximately enough to do the J simple vs. simple tests

$$\mathcal{H}^*_{0,j} : (P, Q) = (P_j, Q_j) \quad \text{vs.} \quad \mathcal{H}_{1,j} : (\tilde{P}_j, \tilde{Q}_j).$$

Now, if we can reject in all J tests under a multiple comparisons procedure we have a Frequentist test of (1). If we can't reject all J nulls, problems remain. Overall, in contrast to (3), Frequentist reasoning is too precious to be disturbed by refutation.

The Frequentist parametric approach will reduce J and so be simpler than the Frequentist nonparametric approach – at the cost of specifying a parametric family. The Bayes parametric approach is likewise simpler than the NPB approach but also has the cost of specifying a parametric family. Neither parametric reduction is persuasive.

Thus, the NPB prescription for finding a virtual control group is to find sets $\mathcal{D}_C$ that let us reject in (1) or (2). This is easier to implement and interpret than a Frequentist analysis and should also give better results – as Bayes tests commonly do.

References

Berger, J. (2003). "Could Fisher, Jerffreys, and Neyman have agreed on testing?" *Statistical Science*, 18: 1–32.

Berger, J. and Bayarri, S. (2004). "The interplay of Bayesian and Frequentist analysis." *Statistical Science*, 19: 58–80.

Borgwardt, K. and Ghahramani, Z. (2009). "Bayesian two-sample tests." URL arXiv:0906.4032[cs.LG]

Chen, C.-F. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society Series B*, 47: 540–546.

Holmes, C., Caron, F., Griffin, J., and Stephens, D. (2012). "Two-sample Bayes non-parametric hypotheis tests." URL arXiv:0910.5060v2[stat.ME]

M. Eaton, A. S., R. Muirhead (2013). "On the limiting behavior of the probability of claiming superiority in a Bayesian context." *Bayesian Analysis*, 8: 221–232.

Tokdar, S., Chakrabarti, A., and Ghosh, J. (2010). "Bayesian nonparametric goodness of fit tests." In Sun, M. C. D. D. P. M. D. and Ye, K. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, chapter 6.1. Springer.

Walker, S. and Muliere, P. (2003). "A bivariate Dirichlet process." *Statistics and Probability Letters*, 64: 1–7.

Comment by Andrew Gelman[6]

Müller and Mitra present an excellent motivation and overview of Bayesian nonparametric models, and in fact their article could have gone on longer, to include models such as Bayesian additive regression trees (Chipman, George, and McCulloch, 2010) which have the potential to revolutionize the practice of causal inference by allowing researchers to directly model potential outcomes (Hill, 2011), avoiding the traditional and often counterproductive focus on average treatment effects and restricted domains of inference. And I am sure there are many other areas of application where Bayesian nonparametrics can allow for scientific advances by allowing researchers to focus on modeling phenomena of interest rather than getting distracted by issues of identification and functional forms.

Bayesian data analysis can be fruitfully considered as an iteration of three steps: (1) model building, (2) inference, and (3) model checking. Compared to traditional Bayesian methods, nonparametric Bayes represents an additional modeling investment in step 1, with the gains coming in step 2 (more accurate models and predictions) and in step 3 (better fit to data). Nonparametric models deserve more attention within Bayesian statistics, and we have added several chapters on them for the upcoming third edition of our book (Gelman et al., 2013).

For all their flexibility, however, nonparametric models are still models. They have assumptions and their fit to data can be checked by comparing observed data to hypothetical replicated datasets simulated from the fitted model (Rubin, 1984, Gelman, Meng, and Stern, 1996). The good news is that, in an environment in which models are fit using posterior simulations, it is typically trivial (in both the mathematical and computational senses) to simulate replicated datasets. Based on our own experiences, we think the most effective model checks are graphicalbut this is no problem either, as such checks are a simple step forward beyond the graphical displays of inferences and data that are becoming standard best practice in nonparametric inference (as illustrated, for example, in Figures 1, 7, and 9 of the paper under discussion). The same sorts of displays that are informative about data can directly be used to explore model fit by comparison to simulated replications.

References

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* 4, 266-298.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, third edition. London: Chapman and Hall.

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733–807.

[6]Department of Statistics, Columbia University, New York, NY

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Comment by Miroslav Kárný[7,8]

The authors have done a great job in describing the state of the art of Bayesian Non-Parametrics and have illustrated the ideas by interesting examples. Their presentation has one (quite wide-spread) methodological flaw I want to point to. Essentially, their paper answers the question "how" and misleads in answering the question "why". They do not take seriously Box's statement they cite (*all models are wrong*). Taking it literally, it would mean that the prior distribution should have support out of the model class (irrespectively of finite or massive parametrisation) and no inference would be possible. Luckily enough, a straightforward inspection of the Bayes rule leads to the Sanov-type view, Sanov (1957); Berec and Kárný (1997), that the posterior distribution is to be interpreted as the probability that a model, within the considered model class, which does not contain reality in the generic case, is the *best projection of reality to this class*. Consequently,

- the non-parametric (massive parametric) inference is susceptible to the same problems as the standard parametrisation (for instance, ignoring continuity of the estimated distribution can cause non-acceptable modelling errors);
- the information about concentration of the posterior distribution is the information regarding how close we are to the best projection and not how close we are to reality: it is increased due to the massive parametrisation but *not* due to better information about closeness to reality;
- the entropy rate, which often reduces to the Kullback-Leibler divergence, is the only adequate Bayes-rule induced measure of closeness.

Technically, the objection against mixture-type modelling is not completely correct as progress in this respect is enormous and counteracts the curse of dimensionality (R. Bellman, Bellman (1961)), which is an inherent barrier of non-parametric inference. Please, take our work Kárný et al. (2006) as an example of a strong research and development stream in this respect.

References

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press, NJ.

Berec, L. and Kárný, M. (1997). "Identification of reality in Bayesian context." In Warwick, K. and Kárný, M. (eds.), *Computer-Intensive Methods in Control and Signal Processing*, 181–193. Birkhäuser.

Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., and Tesař, L. (2006). *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. London: Springer.

[7]Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic

[8]The work was supported by the grant GAČR 13-13502S

Sanov, I. (1957). "On probability of large deviations of random variables." *Matematičeskij Sbornik*, 42: 11–44. (in Russian), also in Selected Translations Mathematical Statistics and Probability, I, 1961, 213–244.

Comment by Michalis Kolossiatis[9]

By writing a condensed, but clear paper, the authors give an overview of Bayesian nonparametric models, as well as cases where these models are particularly useful and (should be) preferred over parametric ones.

The class of Bayesian nonparametric models is, of course, huge. I would like to draw attention to a class of such models which can be used for the joint modelling of several distributions. This class of models, introduced by Griffin et al. (2013), is called Correlated Normalized Random Measures with Independent Increments (CNRMI) and it is constructed by normalising sums of some underlying random measures with independent increments. By using a selection matrix, the modeller can explicitly state which of these underlying measures are shared by which of the distributions, according to the problem in hand. Another interesting feature of this model is that, by using appropriate prior distributions for some of the parameters, a formal model selection can be conducted. Two interesting subclasses are the Correlated Dirichlet Process (CDP) and the Correlated Normalised Generalised Gamma Process (CNGG), where the underlying processes are gamma and generalised gamma and the marginal processes are Dirichlet and normalised generalised gamma, respectively. A similar model, for the case of two distributions, and from a more theoretical perspective, was developed in Lijoi et al. (2013).

In practice, this model will be used in an intermediate part of a larger hierarchical model, as in most BNP models. Simulating from it can be done using standard MCMC methods, for example the slice sampling algorithm for normalized random measure mixture models of Griffin and Walker (2011). This answers, in general terms, the "how" this model can be used. As to "why" it should be used, the reason is the possibility of flexibly and jointly modelling an arbitrary number of correlated distributions, with a direct method of modelling the common parts in any subset of those distributions, and therefore the correlation between the distributions.

Regarding more tangible applications, Griffin et al. (2013) apply the proposed model on survival data and on stochastic frontier data. In the latter case, we have a regression model with two additive error terms. The first set of errors is assigned a normal distribution, whereas the distributions of the second error terms are assumed to follow a CNGG. This model can also be naturally applied in the case of the prostate cancer study data (Example 2 in the paper): the regression on the longitudinal covariate can be applied in a similar fashion as in Zhang et al. (2010), whereas the distributions G_1 and G_2 can follow a CNRMI (or, more specifically, a CDP or a CNGG), without taking into account the possibility of cure. In order to account for the event of cure, the model can be naturally extended by adding a pair of degenerate processes, one for each of the distributions of G_1, G_2 (for example, DPs with base distributions being atomic at t_c, the survival time assigned to cured individuals).

[9]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK

References

Griffin, J. E., Kolossiatis, M., and Steel, M. F. J. (2013). "Comparing distributions by using dependent normalized random-measure mixtures." *Journal of the Royal Statistical Society - Series B*, to appear.

Griffin, J. E. and Walker, S. G. (2011). "Posterior simulation of normalised random measure mixtures." *Journal of Computational and Graphical Statistics*, 20: 241–259.

Lijoi, A., Nipoti, B., and Prüenster, I. (2013). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, to appear.

Zhang, S., Müller, P. M., and Do, K.-A. (2010). "A Bayesian semiparametric survival model with longitudinal markers." *Biometrics*, 66(2): 435–443.

Comment by Athanasios Kottas[10], Maria DeYoreo[10] and Valerie Poynor[10]

We commend the authors for an interesting review of applications of Bayesian nonparametric modeling and inference. Here, we offer some additional discussion, results and references on fully nonparametric regression, which we believe is a key success story of Bayesian nonparametrics.

As the authors discuss in Section 4, two dominant trends in the Bayesian regression literature have been to develop flexible regression function models and to accompany the regression relationship with more comprehensive uncertainty quantification. For problems involving a small to moderate number of random covariates, the *curve fitting regression* approach is an appealing alternative. Specifically, a DP mixture model, $f(y, \boldsymbol{x}; G) = \int k(y, \boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta})$, $G \sim \mathrm{DP}(\alpha, G_0)$, is used for the joint distribution of the response, y, and covariates, $\boldsymbol{x}$, from which inference emerges for the conditional response distribution, $f(y \mid \boldsymbol{x}; G)$. Modeling the joint response-covariate distribution is natural for many applications, especially in the environmental and biomedical sciences.

Although the approach (based on normal mixtures) has been proposed in Müller et al. (1996), it has been overlooked as a general nonparametric regression framework until relatively recently. This may be attributed to the limitations of posterior predictive estimation for full inference about the conditional distribution $f(y \mid \boldsymbol{x}; G)$. However, with posterior simulation extended to the mixing distribution G, the DP mixture curve fitting approach enables rich inference for response densities that can change in non-trivial fashion across the covariate space, and for non-linear regression relationships built from the mean or from percentiles of the response distribution (Taddy and Kottas 2009, 2010). Moreover, the methodology can be extended to handle categorical responses (Shahbaba and Neal 2009; Dunson and Bhattacharya 2011), and looking beyond the standard regression setting, to develop emulation and calibration techniques for stochastic computer simulators (Farah 2011) as well as modeling for marked Poisson processes (Taddy and Kottas 2012).

A particularly promising direction involves problems with (possibly multivariate) ordinal responses, y, which can be represented as discretized versions of latent continuous responses, z, with a DP mixture model employed for the joint distribution of z and $\boldsymbol{x}$. For continuous covariates, the mixture kernel can be built from a multivariate normal which, in the presence of binary responses, requires identifiability restrictions for its covariance matrix (DeYoreo and Kottas 2013). This modeling approach enables flexible nonparametric inference for the implied response classification probabilities, $\Pr(y = j \mid \boldsymbol{x}; G)$, the number of which increases significantly in multivariate ordinal regression problems rendering semiparametric modeling infeasible. Figure 3 illustrates the capacity of the model to uncover both relatively standard and non-monotonic shapes for the ordinal regression relationships as well as non-trivial interactions among covariates.

The curve fitting regression framework can be enhanced with hierarchically depen-

[10]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA

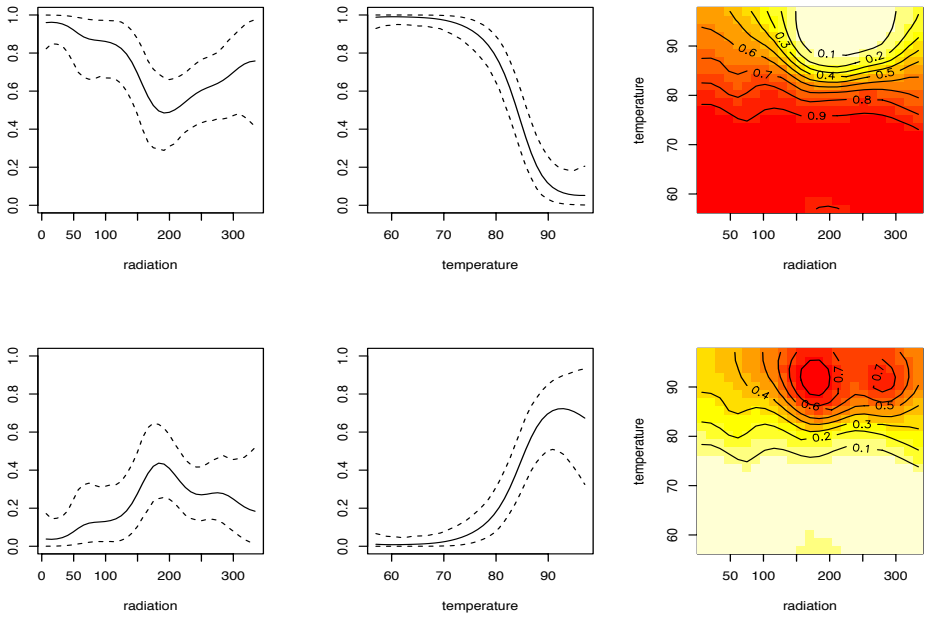

Figure 3: Ordinal regression example using data (available in R) on ozone concentration (variable z measured in ppb), radiation (variable x_1 in langley units) and temperature (variable x_2 in degrees Fahrenheit) recorded over 111 days from May to September of 1973 in New York. Ozone concentration is discretized to construct an ordinal response, y, with classifications of "low", "medium", and "high" concentration corresponding respectively to: $y = 1$ for $z \leq 50$ ppb; $y = 2$ for 50 ppb $< z \leq 100$ ppb; and $y = 3$ for $z > 100$ ppb. Inference results are based on a trivariate normal DP mixture model for (z, x_1, x_2). The top and bottom row panels include inference results for the "low" and "medium" categories; specifically, for $j = 1, 2$, the left and middle columns show posterior mean and 95% interval estimates for $\Pr(y = j \mid x_1; G)$ and $\Pr(y = j \mid x_2; G)$, and the right column plots the posterior mean estimate of $\Pr(y = j \mid x_1, x_2; G)$.

dent nonparametric priors (e.g., Rodriguez et al. 2009). More generally, it can be utilized complementary to DDP regression, with survival analysis providing a practically important area of application. Survival regression problems typically include a treatment categorical factor in addition to random covariates $\boldsymbol{x}$. For instance, for a generic treatment/control setting (indicated by $s \in \{T, C\}$), the joint response-covariate distribution can be modeled with $f(y, \boldsymbol{x}; G_s) = \int k(y, \boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}G_s(\boldsymbol{\theta})$. The choice of the kernel is important to ensure desirable properties for key functions of the survival response distribution, such as the hazard and mean residual life functionals (Poynor and Kottas 2013). Assigning a DDP prior to the pair of mixing distributions (G_C, G_T) results in related regression relationships through the dependent T/C response distributions $f(y \mid \boldsymbol{x}; G_s)$. Here, a variable-weights DDP prior, $G_s = \sum_{h=1}^{\infty} \pi_{sh} \delta_{\tilde{\theta}_h}$, is an attractive

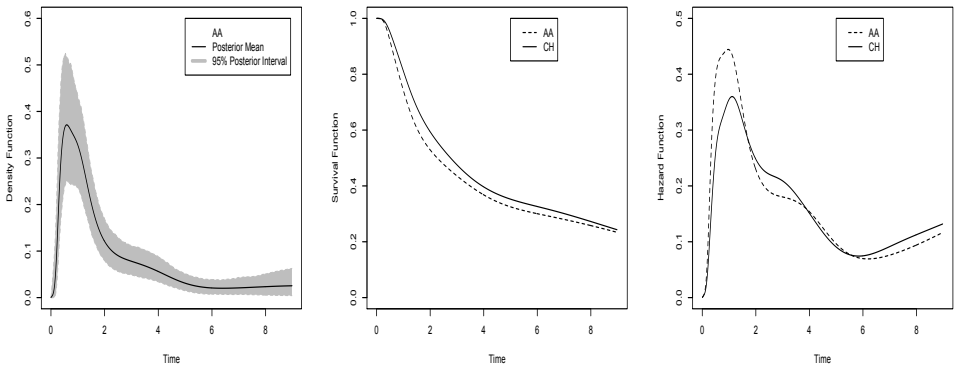

Figure 4: Prostate cancer study (Example 2 of the paper). Inference results are based on a gamma DDP mixture model, $f(y; G_s) = \int \text{gamma}(y; \theta, \phi) \text{d}G_s(\theta, \phi)$, where $s \in \{\text{AA}, \text{CH}\}$, with a variable-weights DDP prior assigned to $(G_{\text{AA}}, G_{\text{CH}})$, using one of the bivariate beta distributions from Nadarajah and Kotz (2005) for the DDP stick-breaking weights. The left panel plots the posterior mean and 95% uncertainty bands for the treatment AA density function. The middle and right panels show the posterior mean estimates under the two treatments for the survival functions and for the hazard rate functions, respectively.

alternative to the basic DDP model (expression (8) of the paper); incorporating dependence through the DDP weights is invariant to the mixture kernel dimensionality, and for this application, it may be more natural to envision similar mixture locations with prevalence varying according to the T/C groups. To retain the DP marginally, we need an appropriate bivariate beta distribution for the latent variables $(\upsilon_{Ch}, \upsilon_{Th})$ that define the stick-breaking weights. For an illustration, Figure 4 shows results based on the portion of the data from the prostate cancer study made available on-line. Note that the point estimates for the hazard functions suggest a non-proportional hazards relationship for TTP under the two treatments providing further demonstration for the practical utility of flexible Bayesian nonparametric modeling relative to traditional parametric or semiparametric regression models.

References

DeYoreo, M. and Kottas, A. (2013). "A fully nonparametric modeling approach to binary regression." Technical Report UCSC-SOE-13-03, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

Dunson, D. B. and Bhattacharya, A. (2011). "Nonparametric Bayes regression and classification through mixtures of product kernels." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M.

(eds.), *Bayesian Statistics 9, Proceedings of the Ninth Valencia International Meeting*, 145–164. Oxford University Press.

Farah, M. (2011). "Bayesian nonparametric methods for emulation, sensitivity analysis, and calibration of computer simulators." Ph.D. thesis, *Statistics and Stochastic Modeling*, University of California, Santa Cruz.

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83: 67–79.

Nadarajah, S. and Kotz, S. (2005). "Some bivariate beta distributions." *Statistics*, 39: 457–466.

Poynor, V. and Kottas, A. (2013). "Nonparametric Bayesian inference for mean residual life functions in survival analysis." Technical Report UCSC-SOE-13-04, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2009). "Bayesian nonparametric functional data analysis through density estimation." *Biometrika*, 96: 149–162.

Shahbaba, B. and Neal, R. (2009). "Nonlinear models using Dirichlet process mixtures." *Journal of Machine Learning Research*, 10: 1820–1850.

Taddy, M. and Kottas, A. (2009). "Markov switching Dirichlet process mixture regression." *Bayesian Analysis*, 4: 793–816.

— (2010). "A Bayesian nonparametric approach to inference for quantile regression." *Journal of Business and Economic Statistics*, 28: 357–369.

— (2012). "Mixture modeling for marked Poisson processes." *Bayesian Analysis*, 7: 335–362.

Comment by Susan M. Paddock[11,12] and Terrance D. Savitsky[11,12]

Müller and Mitra's contribution regarding the practical *whys* and *hows* of non-parametric Bayes (NPB) is welcome. In that spirit, we highlight one basic and one complex social science example for which NPB is uniquely well-suited.

Depression symptoms scores were collected from $n = 299$ clients on three occasions – pre-treatment, post-treatment, and follow-up – during a study of group therapy's effectiveness for treating depression. Clients completed up to four group therapy modules and could join the therapy group at start of a module. Therapy group-induced correlations among client outcomes could thus be modeled using random module effects, which would be linked to post-treatment outcomes via multiple membership, and client-specific growth parameters (*e.g.,* random intercept, time, and quadratic time effects) could be specified for modeling within-client correlations and deviations from the average depression score trajectory (Paddock and Savitsky 2013).

Basic use of NPB for a very common analytic problem. Randomly-sampled depression score trajectories for six clients show convex and concave patterns (Figure 5), so including quadratic time effects in the model seems appropriate. However, for model identifiability, conventional parametric growth curve modeling requires $d + 2$ repeated observations for a polynomial trajectory of degree d (Bollen and Curran 2006). Ad-hoc parameter constraints would thus be required, such as assuming the quadratic time client random effect variance is 0 or setting variance terms equal to a constant (Little et al. 2006), or imposing identifiability through the prior.

Paddock and Savitsky (2013) avoid making ad-hoc constraints by modeling the client growth parameters using a Dirichlet process (DP). The positive probability of ties under DP facilitates a useful parameter dimension reduction, providing a compromise between assuming one trajectory applies equally well to everyone versus having n distinct trajectories for all clients arising from a parametric distribution. There were about 10 unique sets, or clusters, of growth parameters at each MCMC iteration in Paddock and Savitsky (2013). By 'letting the data speak' about which patterns existed in the data, the DP approach captured both convex and concave growth curves, whereas the parametric approach only captured concave curves. DP is particularly promising for such longitudinal intervention studies, considering such typically small numbers of observations per client. Example 6 of Müller and Mitra's paper has similar features - *e.g.,* three random effects and three observations per tripeptide/tissue pair. We would be interested in the authors' comments on whether and how such dimension reduction played a role in the parametric empirical Bayes versus semiparametric comparison.

More complex example. Paddock and Savitsky's (2013) model would constrain module random effects to be constant over time, not allowing for changes in correlations among outcomes for clients who attend modules together. However, client outcomes

[11]RAND Corporation
[12]Supported by NIH/NIAAA Grant R01AA019663

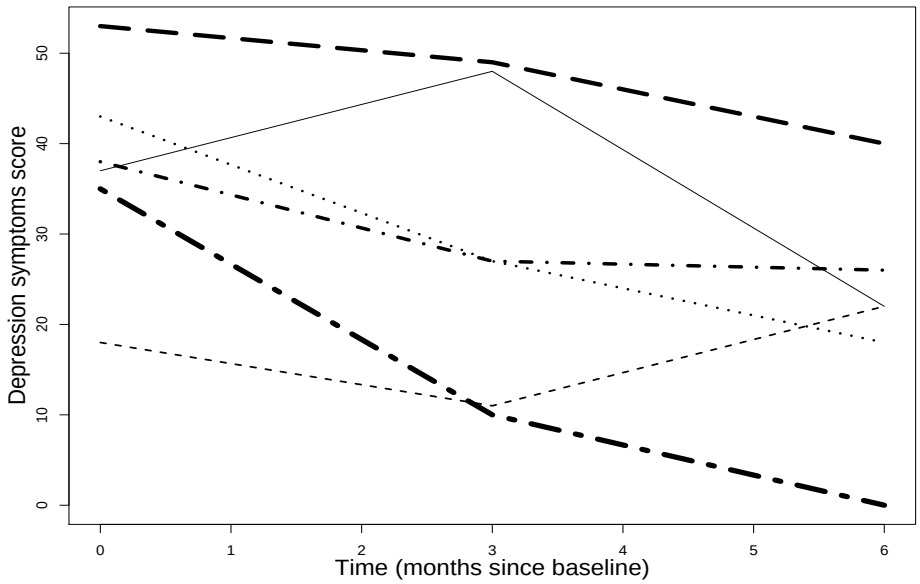

Figure 5: Depression symptoms scores for six randomly-selected clients at $0, 3$, and 6 months post-baseline

might be more strongly correlated at different time points, such as immediately following group therapy versus at baseline or follow-up. Not all clients benefit similarly from group therapy (Smokowski et al. 2001); module effects might change over time and the effects of modules on participant outcome trajectories may vary across study participants. Savitsky and Paddock (to appear)'s dependent Dirichlet process (DDP) model for repeated measures multiple membership data accounts for this and improves model fit. A set of random distributions for client random effect parameters is indexed by therapy group module attendance sequences. Figure illustrates the heterogeneity in the relative effectiveness of group therapy modules. There are clusters of clients whose outcome trajectories vary across modules. Uncovering this variation motivates future research to understand why such variation exists and for whom do module effects vary. Savitsky and Paddock (to appear) found that a parametric additive model alternative for both module and client effects that allowed for time variation failed to capture this heterogeneity.

References

Bollen, K. and Curran, P. (2006). *Latent curve models: A structural equation approach*. Hoboken, NJ: Wiley.

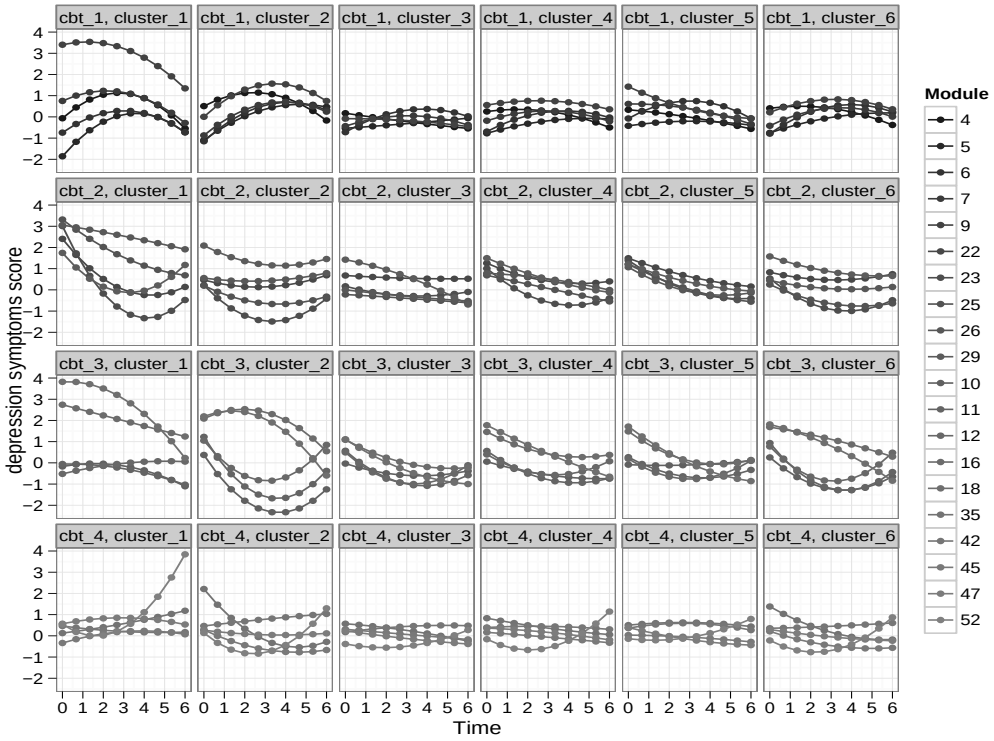

Figure 6: DDP model output. Each curve represents the posterior mean of client depression symptoms score trajectories for a randomly selected module. Each row represents one of four distinct therapy groups in the study ($\text{cbt}_{\text{-}1}, \ldots, \text{cbt}_{\text{-}4}$). The six columns correspond to the largest clusters of clients (from largest to smallest).

Little, T., Bovaird, J., and Slegers, D. (2006). "Methods for the Analysis of Change." In Mroczek, D. and Little, T. (eds.), *Handbook of Personality Development*, 181–211. Mahwah, NJ: Erlbaum.

Paddock, S. M. and Savitsky, T. D. (2013). "Bayesian Hierarchical Semiparametric Modeling of Longitudinal Post-treatment Outcomes from Open-enrollment Therapy Groups." *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 176(3): to appear.

Savitsky, T. and Paddock, S. (to appear). "Bayesian Non-Parametric Hierarchical Modeling for Multiple Membership Data in Grouped Attendance Interventions." *Annals of Applied Statistics*.

Smokowski, P., Rose, S., and Bacallao, M. (2001). "Damaging experiences in therapeutic

groups: How vulnerable consumers become group casualties." *Small Group Research*, 32(2): 223–251.

Comment by G. Parmigiani[13] and L. Trippa[13]

We vividly congratulate Müller and Mitra for elucidating important motivations for the application and theoretical development of Bayesian nonparametric (BNP) models, using illuminating examples. Müller and Mitra make a compelling case for the whys of BNP and illustrate the fascinating ease of computational implementations in many of their examples. We find ourselves in agreement with much of the motivation of this review. A point that may be worth reiterating is that the boundary between parametric and nonparametric approaches is blurred in Bayesian modeling, even more than it is elsewhere. First, many BNP solutions can be alternatively conceptualized as models with very high dimensional parameter spaces. Also, practically, in any given data analysis, parametric and nonparametric elements can coexist, and in fact combinations of parametric structures and simple Bayesian nonparametric tools are common and powerful. These allow one to capture important, and sometime hidden, aspects of the data without sacrificing the parametric models' beauty: thoughtful assumptions, and interpretable estimands. BNP modeling is a natural step when standard parametric tools poorly represent the data; we can then choose to either embed the parametric model into a more flexible BNP construction, or consider competing parametric models by adding parameters; the latter strategy can be time consuming, and challenging from a model selection perspective. BNP modeling versus non standard parametric distributions for the error term in regression problems exemplify the two strategies.

Müller and Mitra emphasize the large spectrum of models and computational procedures developed in recent years. Dirichlet process mixtures and the use of basis functions remain milestones, while several new probabilistic constructions such as recent prior models for relational data Griffiths and Ghahramani (2011); Miller, Griffiths, and Jordan (2009) support Bayesian nonparametric inference in new applied fields. Another aspect that is nicely captured in the review is the analytical tractability of some BNP probabilistic constructions such as Polya Trees, Product partition models, and the more recent Indian Buffet process. In our view, a further motivation to the increasing popularity on BNP tools is in the close connections with established procedures such as k-means clustering Kulis and Jordan (2011), support vector machines and kernel regression methods Schölkopf and Burges (1999). Such connections offer a fertile ground, not only for the interpretation of a number of learning procedure, but also for applications, and for conceiving novel methodologies.

The case studies discussed by Müller and Mitra concisely represent the spectrum of problems and probability models in BNP. Fortunately, this spectrum is wide, and growing wider: the inferential problems and the data types amenable to analysis through BNP modeling is not constrained to any particular structure. We hope that their work will further contribute to yet more creative modeling and a more widespread use of BNP techniques in practice.

[11]Dana Farber Cancer Institute and Harvard School of Public Health, Boston, MA

References

Griffiths, T. L. and Ghahramani, Z. (2011). "The Indian buffet process: An introduction and review." *Journal of Machine Learning Research*, 12: 1185–1224.

Kulis, B. and Jordan, M. I. (2011). "Revisiting k-means: New algorithms via Bayesian nonparametrics." *arXiv preprint arXiv:1111.0352*.

Miller, K., Griffiths, T., and Jordan, M. (2009). "Nonparametric latent feature models for link prediction." *Advances in neural information processing systems*, 22: 1276–1284.

Schölkopf, B. and Burges, C. J. (1999). *Advances in kernel methods: support vector learning*. The MIT press.

Comment by François Perron[14]

The paper is well written. Several applications are discussed, interesting aspects are explored, the review of the literature is good. As a researcher I will require that my students read this paper, amongst others. I find the paper easy to read and the examples are very helpful. Some of the difficulties are presented and classical solutions are given (a.s. discreteness associated with the Dirichlet prior corrected with mixtures, similar things with Polya tree priors when $a_\epsilon = \alpha G^\star(B_\epsilon)$ is corrected by fixing $a_{\epsilon x} = c_m \alpha G^\star(B_{\epsilon x})/G^\star(B_\epsilon), c_m = cm^2$, the fact that it might be necessary to be clever when the time comes to develop partitions in higher dimensions for the Polya tree priors, random partitions, etc.). Representing a function through a Fourier series with a fixed basis and unknown coefficients is interesting and clearly different from the other parts of the paper. For that same reason, I would have liked it if the authors could have talked about the method of sieves in Bayesian statistics. The section on asymptotics is soft. The authors should have said something about the Bernstein-von Mises theorem, see Kleijn and van der Vaart (2012), Rivoirard and Rousseau (2012) for instance. In practice, people will use finite Polya trees priors, the problem of truncation is relevant, the authors should have said something about it. Is the choice of a prior simply a mathematical tool or not? In many situations, Bayesian nonparametric methods require imagination. For example, a copula in dimension d is a probability measure with uniform margins. A direct approach consists of developing a prior on the copula space. An attempt has been made by Dortet-Bernadet (2005) using Polya trees but the flexibility under this method is quite severe. The cumulative distribution function of a random vector can be written in terms of marginal distribution functions and a copula. Sklar's theorem says that if the margins are continous then the copula is unique. The transformation giving the copula C coming from a cdf F is known. Moreover, it is easy to create a prior on the space of the cumulative distribution functions with continuous margins (using the ideas of Hanson for instance, see Hanson et al. (2012)). Therefore, we have an indirect approach inducing a prior on the copula space. In bivariate extreme value theory there is the spectral measure which can be viewed as a probability measure on $[0, 1]$ with mean value $1/2$. Building priors on the space of the cdf on $[0, 1]$ with median $1/2$ is quite easy with Polya tree priors while fixing the mean value at $1/2$ is not. Basically, Bayesian non parametric inference is appealing but it is not always straightforward to implement.

References

Dortet-Bernadet (2005). *Bayesian inference on copulas and tests of independence*. Preprint 2005-10. Department of Statistics, School of Mathematics, University of New South Wales, Sydney, Australia.

Hanson, T. E., Jara, A., and Zhao, L. (2012). "A Bayesian semiparametric temporally-stratified proportional hazards model with spatial frailties." *Bayesian Anal.*, 7(1): 147–188.

[14]Université de Montréal, Montéal, Canada

Kleijn, B. J. K. and van der Vaart, A. W. (2012). "The Bernstein-Von-Mises theorem under misspecification." *Electron. J. Stat.*, 6: 354–381.
URL http://dx.doi.org/10.1214/12-EJS675

Rivoirard, V. and Rousseau, J. (2012). "Posterior concentration rates for infinite dimensional exponential families." *Bayesian Anal.*, 7(2): 311–333.
URL http://dx.doi.org/10.1214/12-BA710

Comment by Christian P. Robert[15] and Judith Rousseau[15]

We congratulate the authors for this very pleasant overview of the type of problems that are currently tackled by Bayesian nonparametric inference and for demonstrating how prolific this field has become. We do share the authors' viewpoint that many Bayesian nonparametric models allow for more flexible modelling than parametric models and thus capture finer details of the data. BNP can be a good alternative to complex parametric models in the sense that the computations are not necessarily more difficult in Bayesian nonparametric models. However we would like to mitigate the enthusiasm of the authors since, although we believe that Bayesian nonparametric inference has proved extremely useful and interesting, we think they oversell the "nonparametric side of the Force"! Our main point is that by definition, Bayesian nonparametric inference is based on prior probabilities that live on infinite dimensional spaces and thus are never completely swamped by the data. It is therefore crucial to understand which (or why!) aspects of the model are strongly influenced by the prior and how.

As an illustration, when looking at Example 1 with the censored zeroth cell, our reaction is that this is a problem with no proper solution, because it is lacking too much information. In other words, unless some parametric structure of the model is known, in which case the zeroth cell is related with the other cells, we see no way to infer about the size of this cell. The outcome produced by the authors is therefore unconvincing to us in that it seems to only reflect upon the prior modelling $(\alpha, G^\star)$ and not upon the information contained in the data. Now, this prior modelling may be to some extent justified based on side information about the medical phenomenon under study, however its impact on the resulting inference is palpable.

Recently (and even less recently) a few theoretical results have pointed out this very issue. E.g., Diaconis and Freedman (1986) showed that some priors could surprisingly lead to inconsistent posteriors, even though it was later shown that many priors lead to consistent posteriors and often even to optimal asymptotic frequentist estimators, see for instance van der Vaart and van Zanten (2009) and Kruijer et al. (2010). The worry about Bayesian nonparametrics truly appeared when considering (1) asymptotic frequentist properties of semi-parametric procedures; and (2) interpretation of inferential aspects of Bayesian nonparametric procedures. It was shown in various instances that some nonparametric priors which behaved very nicely for the estimation of the whole parameter could have disturbingly suboptimal behaviour for some specific functionals of interest, see for instance Arbel et al. (2013) and Rivoirard and Rousseau (2012). We do not claim here that asymptotics is the answer to everything however bad asymptotic behaviour shows that something wrong is going on and this helps understanding the impact of the prior. These disturbing *bad results* are an illustration that in these infinite dimensional models the impact of the prior modelling is difficult to evaluate and that although the prior *looks* very flexible it can in fact be highly *informative* and/or restrictive for some aspects of the parameter. It would thus be wrong to conclude that every aspect of the parameter is well-recovered because some are. This has been a well-known fact for Bayesian parametric models, leading to extensive research on reference

[15]Université Paris-Dauphine, Paris, France

and other types of objective priors. It is even more crucial in the nonparametric world. No (nonparametric) prior can be well-suited for every inferential aspect and it is important to understand which aspects of the parameter are well-recovered and which ones are not.

We also concur with the authors that Dirichlet mixture priors provide natural clustering mechanisms, but one may question the "natural" label as the resulting clustering is quite unstructured, growing in the number of clusters as the number of observations increases and not incorporating any prior constraint on the "definition" of a cluster, except the one implicit and well-hidden behind the non-parametric prior. In short, it is delicate to assess what is eventually estimated by these clustering methods.

These remarks are not to be taken as criticisms of the overall Bayesian nonparametric approach, just the contrary. We simply emphasize (or recall) that there is no such thing as a free lunch and that we need to post the price to pay for potential customers. In these models, this is far from easy and just as far from being completed.

References

Arbel, J., Gayraud, G., and Rousseau, J. (2013). "Bayesian adaptive optimal estimation using a sieve prior." *Scandinavian Journal of Statistics*, to appear.

Diaconis, P. and Freedman, D. (1986). "On the consistency of Bayes estimates." *Annals of Statistics*, 14: 1–26.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). "Adaptive Bayesian density estimation with location-scale mixtures." *Electronic Journal of Statistics*, 4: 1225–1257.

Rivoirard, V. and Rousseau, J. (2012). "Bernstein von Mises theorem for linear functionals of the density." *Annals of Statistics*, 40: 1489–1523.

van der Vaart, A. and van Zanten, J. H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth." *Annals of Statistics*, 37: 2655–2675.

Comment by James G. Scott[16]

I wish both to congratulate and to thank Drs. Müller and Mitra for their excellent review article. This area moves fast—especially the "how" part of it!—and it is frankly hard for us non-experts to keep up with. In making that task a bit easier, the authors have done us a great service.

Nonetheless, this paper provides an occasion to ring a note of skepticism regarding a very common practice: the use of nonparametric Bayesian models to make statements about clusters and cluster membership. Clustering, like any Bayesian model-selection problem, is very hard. Yet I have heard it glibly asserted—certainly not by Müller and Mitra, but by others who shall remain nameless—that Bayesian nonparametric methods "solve the problem." Surely I am not the only reader of *Bayesian Analysis* to have encountered this claim, nor am I the only one to view it with circumspection.

Müller and Mitra certainly do not endorse the use of BNP methods for answering the "how many clusters" question. They argue that the mere "side effect of creating a random partition" is insufficient to support inference about clusters, as "the prior $p(\rho_n)$ [over number of clusters] includes several often inappropriate features" under the commonly used models. They go on to cite the implicit geometric decay in cluster size as one such inappropriate feature. I wish to draw attention to another: namely, that nonparametric Bayesian priors are not, and cannot be, predictively matched across mixture models of differing size. This fact contra-indicates their use as an "assumption-free" clustering tool.

The notion of predictive matching is simple and appealing, especially in nested-model cases like the clustering problem. One requires p observations $\mathbf{y} = (y_1, \ldots, y_p)$ to identify a p-dimensional sampling model $\mathcal{M}_1$ having parameter space Θ_1. Imagine starting with no prior knowledge and observing such a $\mathbf{y}$, called a minimal training sample. Having used the information in $\mathbf{y}$ to (only just!) identify the larger model, no degrees of freedom remain for comparing $\mathcal{M}_1$ with some smaller nested model $\mathcal{M}_0$ having parameter space $\Theta_0 \subset \Theta_1$. Therefore, for any $\mathbf{y}$ of size p, it should be the case that the two marginal likelihoods are equal:

$$\int_{\Theta_1} p(\mathbf{y} \mid \theta_1)\, d\Pi(\theta_1) = \int_{\Theta_0} p(\mathbf{y} \mid \theta_0)\, d\Pi(\theta_0)\,.$$

If $\Pi(\Theta_1)$ and $\Pi(\Theta_0)$ are such that this holds for all minimal training samples, they are said to be *predictively matched* (see, e.g. Berger and Pericchi 2001). If they are not, then we are forced to claim that one model seems better in light of $\mathbf{y}$. Yet where can this information have come from, given that we have only just identified the parameters of the larger model? Assuming the two models have equal prior probability, the information must have been in $\Pi(\theta_0)$, $\Pi(\theta_1)$, or both—distributions about which we claimed to know nothing at all.

Using predictively mismatched priors for model selection is not wrong *per se*. But

[16]University of Texas at Austin, McCombs School of Business and Division of Statistics and Scientific Computing, Austin, TX

doing so does contribute a bias to the Bayes factor for comparing $\mathcal{M}_1$ and $\mathcal{M}_0$. By "bias", I mean information that arises not from data but from priors over model-specific parameters. This bias is desirable when it reflects real prior belief. It is undesirable when it is an artifact of convenience.

Now to the Dirichlet process, where a simple example illustrates the problem. Data y arrives from a univariate distribution G. We believe that G is a discrete mixture, and therefore model it with a DPM of simple parametric forms:

$$(y_i \mid \theta_i) \sim p(\theta_i)\,, \quad \theta_i \sim G\,, \quad G \sim DP(\alpha, G_0)\,.$$

Under this model, the marginal distribution for the singleton observation y_1 is the same as the marginal distribution for y_1 under the base measure G_0, regardless of the total mass parameter α. Said another way, the marginal likelihood $p(y_1 \mid \alpha)$ is flat as a function of α. The corollary is that, for even two observations, the marginal likelihood $p(y_1, y_2 \mid \alpha)$ is *not* flat in α (see, e.g. Basu and Chib 2003).

But of course it is α that controls the number of mixture components in the model. We therefore conclude that the Dirichlet process mixture model is predictively matched–in the sense of giving the same posterior probabilities across different clustering parameters–to a single observation! For the purpose of model selection, this is startlingly inappropriate. It implies that, after having seen only two observations, one must state a definite opinion about whether one mixture component or two is more likely. Upon what available information can such an opinion rest?

The larger problem, of course, is that predictive matching cannot hold across an infinite family of discrete mixture models. If one simply must entertain all such models, there is no choice but to encode, via the joint prior over the base measure and α, a definite set of beliefs regarding the number and sizes of clusters one expects to see in the data.

In many applications of nonparametric Bayes that I have encountered, it appears customary to obscure these beliefs, but surely impossible to avoid them. The work of Müller and Mitra is a notable exception to this, and would serve as an example worthy of emulation.

References

Basu, S. and Chib, S. (2003). "Marginal likelihood and Bayes factors for Dirichlet process mixture models." *Journal of the American Statistical Association*, 98: 224–235.

Berger, J. and Pericchi, L. (2001). "Objective Bayesian methods for model selection: introduction and comparison." In *Model Selection volume 38 of Institute of Mathematical Statistics Lecture Notes – Monograph Series*, 135–207. Beachwood.

Comment by Surya T. Tokdar[17]

The authors have done a commendable job of showcasing models and examples that champion the cause of Bayesian Nonparametrics for applied statistics. However, as to be expected with any review of a rapidly growing discipline, some gaps remain. This discussion attempts to fill one such gap: *estimating non-crossing quantile curves* for which Bayesian Nonparametrics has provided practical solutions to some 35 year old problems.

Quantile curve estimation is a regression technique where efforts are spent on directly modeling and estimating conditional quantiles. For a proportion $\tau \in (0,1)$ let $Q_Y(\tau|X) = \inf\{a : P(Y \le a|X) \ge \tau\}$ denote the τ-th conditional quantile of a response Y given a predictor vector X. Koenker and Bassett (1978) introduced the linear quantile regression model: $Q_Y(\tau|X) = \beta_0(\tau) + X^T\beta(\tau)$. "Linear" only refers to linearity in the model parameters, X may include non-linear and interaction terms of the original covariates. The intercept and slope parameters are easily estimated by linear programming and the estimates are consistent, asymptotically Gaussian and robust against outliers. Current literature on quantile regression (QR) is both deep and diverse, with wide ranging applications in economics, public health, ecology, etc.; see Koenker (2005) for a comprehensive overview.

Most scientific applications of QR require inference over a dense grid of τ values, which is usually done by assimilating inference from single-τ model fits (e.g., Elsner et al. 2008). Figure 7(a) shows estimated conditional quantile curves for several τ values for the well-known motorcycle data with $Y =$ "head acceleration" and $X =$ B-splines (df = 15) transforms of "time from impact". The estimates do a great job of capturing heteroskedasticity, i.e., quick changes in the response distribution's shape and spread against time. But a closer look reveals several issues: the curves cross each other violating laws of probability; the waviness and the local optima of the curves change wildly across τ reflecting poor borrowing of information; all quantile curves nearly collapse to a single point at the boundary, where uncertainty should have been high due to data scarcity. Rearrangement (Chernozhukov et al. 2011) avoids the embarrassing issue of crossing (Figure 7(b)), but the other two problems persist.

Reich et al. (2011) and Tokdar and Kadane (2012) use Bayesian Nonparametrics to provide two practicable solutions to jointly estimating non-crossing quantile curves. Both solutions are based on non-parametric priors (a Bernstein basis polynomial prior and a transformed Gaussian process prior, respectively) on the function valued parameters $\beta_0(\tau), \beta(\tau)$, $\tau \in (0,1)$ constrained to $\dot\beta_0(\tau) + x^T\dot\beta(\tau) > 0$ for all $\tau \in (0,1)$ and all x in a pre-specified predictor space. Figure 7(c) shows quantile curves for the motorcycle data estimated with the Tokdar and Kadane approach. The problems of quantile crossing, poor borrowing of information and boundary collapsing are all gone!

Non-crossing quantile curves could also be derived from an estimate of the conditional density of Y given X. However, existing conditional density estimation techniques struggle against heterosckedasticity. Figure 7(d) illustrates this for the Linear

[17]Department of Statistical Science, Duke University, Durham , NC

DDP method (De Iorio et al. 2004). The logistic GP approach of Tokdar et al. (2010) gives similar results. Bayesian nonparametric QR provides better quality estimates at a fraction of the computing cost.

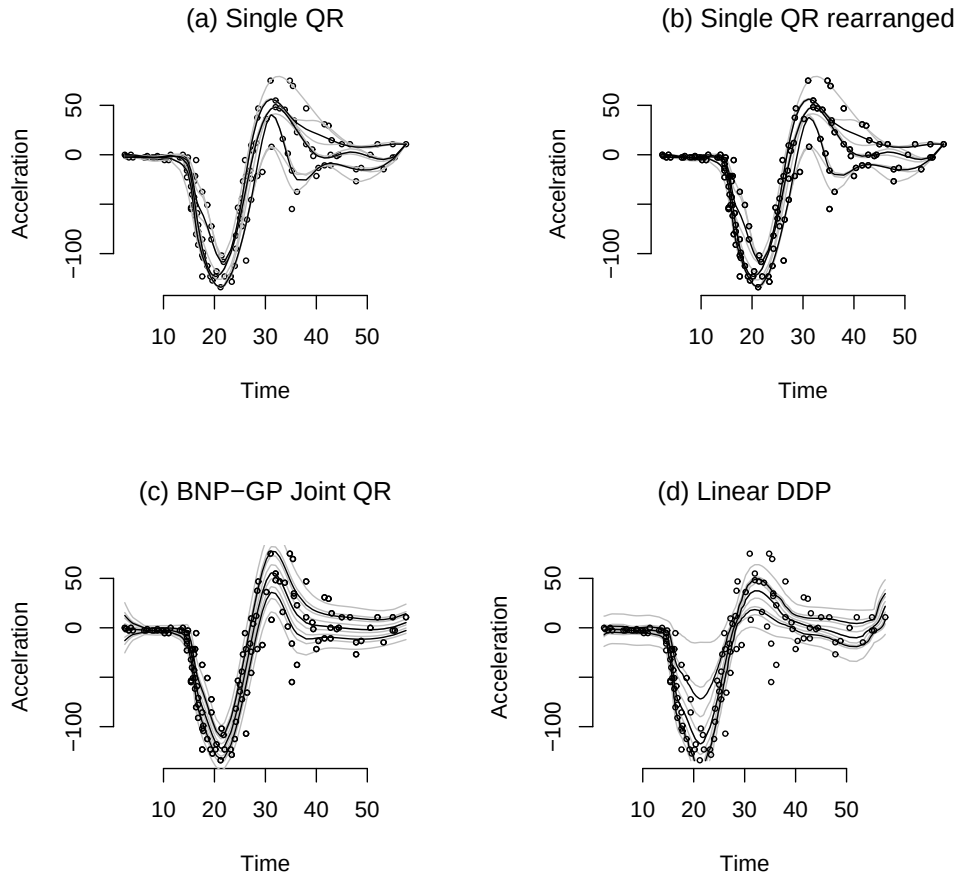

Figure 7: Estimated quantile curves at $\tau \in \{0.1, 0.2, \cdots, 0.9\}$ (gray lines) and at $\tau \in \{0.25, 0.5, 0.75\}$ (black lines) for motorcycle data (open circles). Single QR fits were done with the `rqss()` function of the `quantreg` R-package. Rearrangement was done by obtaining single QR fits over the dense grid $\tau \in \{0.01, 0.02, \cdots, 0.99\}$. Joint QR fits with the transformed GP method of Tokdar and Kadane (2012) was implemented on the same dense grid with codes available at `http://www.stat.duke.edu/~st118/Software/`. Linear DDP was implemented with the R-package `DPpackage` of Jara et al. (2011).

References

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2011). "Quantile and probability curves without crossing." *Econometrica*, 78: 1093–1125.

De Iorio, M., Müller, P., Rosner, G., and MacEachern, S. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215.

Elsner, J. B., Kossin, J. P., and Jagger, T. H. (2008). "The increasing intensity of the strongest tropical cyclones." *Nature*, 455: 92–95.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). "DPpackage: Bayesian Semi- and Non-parametric modeling in R." *Journal of Statistical Software*, 40: 1–30.

Koenker, R. (2005). *Quantile regression, Econometric Society Monographs*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). "Regression quantiles." *Econometrica*, 46: 33–50.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). "Bayesian spatial quantile regression." *Journal of the American Statistical Association*, 106: 6–20.

Tokdar, S. T. and Kadane, J. B. (2012). "Simultaneous linear quantile regression: a semiparametric Bayesian approach." *Bayesian Analysis*, 7: 51–72.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010). "Density regression with logistic Gaussian process priors and subspace projection." *Bayesian Analayis*, 5: 316–344.