Bayesian Analysis (2013) **8**, Number 2, pp. 319–322

Comment on Article by Müller and Mitra

Anthony O'Hagan*

Müller and Mitra have given us a superb paper, eloquently arguing for the many applications of Bayesian nonparametric (BNP) methods. There is no doubt that such techniques have enormous value in a wide range of contexts. I want to raise two linked areas of quite general concern, but these should not be read as detracting in any way from the quality and importance of this paper.

1 Prior information

Bayesian methods require a prior distribution that encodes genuine prior information about the model parameters. I find it quite depressing how rarely this fact is taken seriously in published work which professes to be Bayesian. To illustrate ideas I will look at the prior distribution in the authors' Example 1 and ask what genuine prior information it encodes.

All BNP methods involve specifying a prior distribution for a function. In Example 1, the unknown function in question is the probability mass function F, where $F(y)$ is the probability that a given type of T-cell will be observed y times in the probe. The problem requires that we specify our prior knowledge about F; that is, we need to specify a joint prior distribution for $\{F(0), F(1), F(2), \ldots\}$. This is an infinite-dimensional distribution (as will invariably be the case in BNP applications), so we are looking at a complex problem.

Complex problems benefit from being build up in stages, so first consider a simple parametric model. A natural choice in this problem is to suppose that F is a Poisson distribution $Po(\lambda)$, for some λ, and then to put a prior distribution on λ. The use of the word 'model' here is enlightening — all models involve some degree of simplification of reality. In this case, the parametric model is a simplification of our real prior beliefs. It states that the prior distribution for F gives zero prior probability for all possible distributions F that are not Poisson distributions. What the prior for λ says about F depends to some extent on what judgements were actually used to derive it. Typically, because λ is generally seen as the mean of the Poisson distribution (although of course it is also the variance), its prior distribution will be elicited by making judgements about the mean $\mu(F) = \sum_{y=0}^{\infty} yF(y)$ of F. A few specific judgements such as median and quartiles of $\mu(F)$ will have been made and a convenient distribution fitted to those judgements. The full distribution for F is then completed by a judgement that F is likely to be unimodal and similar to a Poisson distribution, and the parametric assumption of a Poisson distribution is then a convenient choice that is 'fitted' to this judgement.

The underlying approach applies to all pragmatic prior distribution specification: a

*Department of Probability and Statistics, University of Sheffield, Sheffield, UK, a.ohagan@sheffield.ac.uk

DOI:10.1214/13-BA811C

few specific judgements are made or elicited about that distribution, and then the rest is filled in by convenient choices which 'fit' those judgements. We see that approach here in both the specification of a univariate distribution for $\mu(F)$ and then the specification of a prior distribution for all other aspects of F. The simplification of reality lies basically in the second step of making arbitrary, convenient choices to fit the few real, firm judgements that we are prepared to make. Given a particular problem, there are always different ways to simplify, leading to different models, or in this case different prior distribution choices. When eliciting a distribution we can arrive at different distributions in two ways, by starting with different elicited judgements or by having the same judgements but choosing a different distribution to fit them. So now consider alternative prior formulations for the authors' Example 1.

The first kind of BNP model that we might use in this case is a Dirichlet process (DP). Although the authors go straight to a DP mixture model in Example 1, and in the main text argue that this avoids the discreteness of the DP, in Example 1 F is necessarily discrete anyway. So in their notation we could let $F \mid \lambda \sim DP(\alpha, Po(\lambda))$, so that the expectation is a Poisson distribution, and it is natural then to give λ the same prior distribution as in the parametric model. Or is it?

In the parametric model the distribution of $\mu(F)$ in the infinite-dimensional prior distribution for F is the same as the distribution assigned to λ but this no longer holds in the DP model. Conditional on λ, $\mu(F)$ is a random variable. Its expectation is the expectation of λ but there is uncertainty around this value. A formula for the variance can be derived, but is a complicated expression involving α and the prior distribution of λ. I do not think a closed form expression exists for the implied distribution of $\mu(F)$, and therefore we cannot readily fit the DP model to elicited judgements about $\mu(F)$ such as median and quartiles. This is not a problem unique to BNP models. In hierarchical models generally it is not easy to formulate prior distributions for hyperparameters, and the fitting of hierarchical models to judgements about quantities such as $\mu(F)$ that are meaningful in the original problem is rarely addressed properly in the literature; indeed it is usually ignored.

The DP model clearly relaxes the parametric model's very strong assumption that F is exactly Poisson, but any such relaxation requires additional judgements in order to fit the more complex model. In this case, we need something to identify a suitable value for α. This parameter controls how close F is to a Poisson distribution, but it is not easy to see how this might be linked to realistic judgements about F.

The DP mixture model actually employed in Example 1 is given in equation (3) in the paper. Again, although it might be thought natural to equate the distribution G^* to that of λ in the parametric model this is wrong — G^* is not the distribution of $\mu(F)$ in this model.

2 Extrapolation

Here is something else that I think is wrong: statisticians routinely act as if inferences about the parameters of empirically fitted statistical models were meaningful. To illustrate my point, consider a regression problem with one explanatory variable x and a response variable Y. Let the regression function be $\eta(x) = E(Y \mid x)$ and suppose that we know $\eta(0) = 0$. Then a simple parametric model for η might be the linear regression $\eta(x) = \beta x$ with a known distribution of Y around its mean (perhaps with unknown variance σ^2). We proceed to put a prior distribution on β (and if necessary on σ^2) and to derive a posterior distribution for β. But all models are wrong and in reality the regression line will not be linear. In this case, my previous comment would ask, what does the prior distribution of β mean (and how might we specify/elicit it)? We can now also add, what does the posterior distribution of β mean?

These questions raise issues of extrapolation. We know well enough that even if the straight line fits well within the range of the x values in the data it is dangerous to extrapolate in the sense of predicting outside the range of the data. There is, however, another important kind of extrapolation, which is to treat inferences about parameters as meaningful.

Science progresses by a process of extrapolating what has been learnt from individual experiments to give insight into what we might see in other experiments and in real-world situations. Parameters that represent real physical quantities, and which thereby relate to scientific theories, are the key to this kind of extrapolation. When a scientist runs an experiment he or she is doing so in order to learn something of scientific value. That is rarely confined to the specific context of the experiment itself; the scientist rarely wants only to be able to predict the outcome of more repetitions of exactly the same kind of circumstances. No, he or she is generally trying to learn something that can be extrapolated. They are hoping that the statistician analysing those data will provide useful inferences about those physical parameter values.

But the model is wrong. As I said before, the essence of a model is that it simplifies — the true regression relationship may be close to linear but not exactly so, and so on. Even a model that empirically fits the data very well is wrong, and we cannot extract from it reliable inferences about physically meaningful parameters no matter how many observations we have. Consider again the linear regression model. If we get more and more data for x values in a given range, the posterior distribution of β will converge to what is in a well-defined sense (see Walker, 2013) a best-fitting value. But when the model is wrong this limit depends not only on the range of x values over which we are fitting but on the distribution of x within that range. We have all been taught that the x values are ancillary, but that only applies when the model is true. The value of β to which the posterior converges is clearly not of scientific interest in itself and does not extrapolate in either of the above senses.

What about BNP models? A nonparametric regression model of the kind discussed in Section 4 of the paper, particularly the fully nonparametric models in Section 4.3, may not be wrong in the sense of giving zero prior probability in the neighbourhood of the true regression function, but their parameter estimates are no more guaranteed to have scientific meaning than parametric models. They allow us to reliably predict Y within the range of the data (arbitrarily accurately as the sample size increases), which is one way in which they improve on parametric models, but they do not provide a basis for extrapolation. Indeed, they often bury what might be potentially meaningful parameters deeper in the modelling hierarchy.

Statisticians are good at fitting data in such a way as to produce useful predictions of the process in repetitions of the same kind of conditions. But we often cloak those predictions in inferences about model parameters and pretend that these are somehow meaningful. It takes more than that to do real science. It is necessary not only to use nonparametric models but to provide carefully thought out prior information. That is the link between my two apparently disparate comments on Müller and Mitra's paper. I am aware that I have only sketched this link here. More can be found in Brynjarsdottír and O'Hagan (2013), but this is also work in progress.

References

Brynjarsdottír, J. and O'Hagan, A. (2013). Learning about physical parameters: The importance of model discrepancy. Submitted to *SIAM/ASA Journal of Uncertainty Quantification*.

Walker, S. G. (2013). Bayesian inference with misspecified models (with discussion). *Journal of Statistical Planning and Inference* (in press).