# Quantifying the price of uncertainty in Bayesian models

Milovan Krnjajić
*National University of Ireland, Galway*
David Draper
*University of California at Santa Cruz*
e-mail: milovan.krnjajic@nuigalway.ie

**Abstract**

During the exploratory phase of a typical statistical analysis it is natural to look at the data in order to narrow down the scope of the subsequent steps, mainly by selecting a set of families of candidate models (parametric, for example). One needs to exercise caution when using the same data to assess the parameters of a specific model and deciding how to search the model space, in order not to underestimate the overall uncertainty, which usually occurs by failing to account for the second order randomness involved in exploring the modelling space. In order to rank the models based on their fit or predictive performance we use practical tools such as Bayes factors, log-scores and deviance information criterion. Price for model uncertainty can be paid automatically when using Bayesian nonparametric (BNP) specification, by adopting weak priors on the (functional) space of possible models, or in a version of cross validation, where only a part of the observed sample is used to fit and validate the model, whereas the assessment of the calibration of the overall modelling process is based on the as-yet unused part of the data set. It is interesting to see if we can determine how much data needs to be set aside for calibration in order to obtain an assessment of uncertainty approximately equivalent to that of the BNP approach.

***Keywords: model uncertainty, Bayesian non-parametric specification, cross validation, model choice***

## Introduction

When faced with a task of analyzing a data set, statisticians usually take a standard, data-analytic (DA), approach to model specification. In DA approach we explore the space of models in search for the 'right' model, using all of the available data and then using the same data to draw inferential or predictive conclusions conditional on the results of the search. This amounts to using the data twice and often yields poorly calibrated (too narrow) predictive intervals.

There seem to be only two principled solutions to this problem: (1) Bayesian nonparametric (BNP) modelling (with enough data) in which prior distributions are specified on the entire model space, therefore avoiding some of the search and the use of data to specify error distributions, response surfaces, etc., and (2) A version of Bayesian cross-validation (we call it 3-way out-of-sample predictive cross-validation, or 3CV, in a manner somewhat related to a method used in machine learning; 3CV is a modification of DA search in which the data are partitioned into 3 subsets $(S_1; S_2; S_3)$, rather than the usual 2, and where a DA search is undertaken iteratively, modeling with $S_1$ and predictively validating with $S_2$; $S_3$ is not used in quoting final uncertainty assessments, but is instead used to evaluate predictive calibration of the entire modeling process. It looks as if the approach (2) resolves the problem by paying the "right" price for shopping around in the modelling space in terms of setting aside a part of the data.

BNP modeling is often characterized as providing "insurance" against mis-specified parametric models for the following reason: (a) You can generate data from a known ("true")

parametric model $M_1$ and fit $M_1$ and a BNP model to the simulated data sets; both will be valid (both will reconstruct the right answer averaging across simulation replications) but the BNP uncertainty bands will typically be wider. (b) You can also generate data from a different model $M_2$ and fit $M_1$ and BNP to the simulated data sets; often now only BNP will be valid. People refer to the wider uncertainty bands for BNP in (a) as the "insurance premium" you have to pay with BNP to get the extra validity of BNP in (b). But this is not a fair comparison: the simulation results in (a) and (b) were all conditional on a known "true" model, and don't immediately apply to a real-world setting in which you don't know what the "true" model is. However, when you pay an appropriate price for shopping around for the "right" parametric model (as in 3CV), the discrepancy between the parametric and BNP uncertainty bands vanishes.

The approach described above begs the question – can we quantify in a general way (and how exactly) the price of model uncertainty? One idea involves a comparison of *how much data* Bayesian parametric and nonparametric models need *to achieve the same inferential accuracy* about the main quantity of interest. In order to quantify the price of model uncertainty we may proceed as follows: specify a BNP model centered at an a priori plausible parametric model using all $n$ data values and perform the inference; then find out how many data points $n_{DA} < n$ are needed by the best parametric model, discovered with a DA search, to achieve the same inferential accuracy as the BNP model; the difference $(n - n_{DA})$ is how much data should be reserved in 3CV subset $S_3$.

The plan of the paper is as follows: in the first section we describe the simulation setup with a parametric (Poisson based) model and its BNP counterpart and explain differences in estimated inferential and predictive uncertainty. Section 2 describes an attempt to gauge out what fraction of the data set should be used in the calibration stage of a DA model that results in an assessment of uncertainty approximately equivalent to that of the BNP approach.

# 1 Bayesian parametric Poisson based model vs. a BNP model

Assume that we have a data set (of size $n$) consisting of counts coming from an unknown data generating mechanism. The first thing to try parametrically with count data is usually a fixed-effects Poisson (FEP) model (for $i = 1, \ldots, n$):

$$\begin{array}{rcc} (y_i|\theta) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta)] \\ (\theta|\mu,\sigma^2) & \overset{\text{iid}}{\sim} & N(\mu,\sigma^2) \\ (\mu,\sigma^2) & \sim & p(\mu,\sigma^2). \end{array} \tag{1}$$

This specification uses a Lognormal prior for $\lambda = e^{\theta}$ rather than conjugate Gamma choice; the two families are similar, and the Lognormal generalizes more readily. In practice data often exhibit heterogeneity resulting in (extra-Poisson variability), manifesting as variance-to-mean ratio, $VTMR > 1$. A natural parametric extension to FEP would be to try a random effects Poisson model (REP):

$$\begin{array}{rcc} (y_i|\theta_i) & \overset{\text{ind}}{\sim} & \text{Poisson}[\exp(\theta_i)] \\ (\theta_i|G) & \overset{\text{iid}}{\sim} & G \\ G & \equiv & \text{N}(\mu,\sigma^2) \\ (\mu,\sigma^2) & \sim & p(\mu,\sigma^2), \end{array} \tag{2}$$

Here, $i = 1, ..., n$, and we assume a cumulative distribution function (CDF) of latent variables (random effects), $\theta_i$, to be parametric (Gaussian).

The problem is that the mixing distribution, $G$, in the population to which it is appropriate to generalize may be multimodal or skewed, which a single Gaussian can't capture. If so, this REP model can fail to be valid. Moreover, this would usually be diagnosed with something like a density trace of posterior means of $\theta_i$, looking for need to use mixture of Gaussians instead of single one, but choosing $G$ to be Gaussian will tend to make diagnostics support Gaussian model even when it's not right.

Therefore, it would be good to remove the assumption of a specific parametric family (Gaussian) for the mixing distribution $G$ of the random effects, by allowing $G$ to be random and specifying a prior model on the space of $\{G\}$. This BNP model may be centered on a Gaussian model, $N(\mu, \sigma^2)$, but would permit adaptation/learning. Specifying a prior for an unknown distribution requires a stochastic process with realizations (sample paths) that are CDFs. We use the Dirichlet process (DP): $G \sim DP(\alpha\, G_0)$, where $G_0$ is the center or base distribution of the process and $\alpha$ is a precision parameter, see Ferguson (1974). DP mixture Poison model (DPMP: this paper's BNP model):

$$y_i \mid G \overset{ind}{\sim} \int \text{Poisson}(y_i; e^{\theta}) \mathrm{d}G(\theta), \tag{3}$$

where $G$ is a random mixing distribution. For a data set $y = (y_1, ..., y_n)$ the BNP model is:

$$\begin{array}{rcl} y_i \mid \theta_i & \overset{ind}{\sim} & \text{Poisson}(e^{\theta_i}) \\ \theta_i \mid G & \overset{iid}{\sim} & G \\ G & \sim & \text{DP}(\alpha, G_0(\psi)), \end{array} \tag{4}$$

where $\psi = (\mu, \sigma^2), G_0 \equiv N(\cdot; \mu, \sigma^2)$ and $i = 1, ..., n$. Additional model stages are introduced by placing priors on $\alpha$ and $\psi$. MCMC implemented for a marginalized version of DP mixture. Key idea: G is integrated out over its DP prior, resulting in a marginalized version of (4) that follows Pólya urn structure, as shown in Blackwell and MacQueen (1973).

Further references and details of DP mixture modelling along with the description of the simulations with a number of data sets can be found in Krnjajić et al. (2008). Here, it suffices to say that the sample sizes were $n = 300$ and that the data sets were generated based on a variety of unimodal (symmetric and skewed) and bimodal distributions of latent variables (random effects) resulting in data samples with increased variability, nontrivial tails, and densities which were unimodal or with a slight to a noticeable bimodality.

Figure 1 shows the posterior predictive distributions obtained from the parametric REP model and a BNP model with a DP prior, where the data set of counts was generated by a model with a bimodal distribution of latent variables (random effects). (The posterior predictive distribution is always obtained as $p(y^*|y) = \int_\Theta p(y^*|\theta)p(\theta|y)d\theta$) It is obvious from the graphs that the REP model can't adapt to bimodality or skewness without remodelling (say) as a mixture of Gaussians on the latent scale, whereas the BNP modelling smoothly adapts to the data-generating mechanism. A formal comparison of the parametric and the BNP model (using log-scores and deviance information criterion, DIC) showed clear preference for the BNP model when the data were generated with non-Gaussian distribution of random effects.

It is interesting to analyze what is happening on the scale of latent variables which come from random mixing distribution $G$. We can do this since the BNP model permits obtaining posterior draws of $G$, $P(G \mid \text{ data })$. Based on these draws we can compute estimates such as the mean functional, $E[y \mid G)]$, and in fact, obtain the entire distribution of $E[y \mid G)]$.

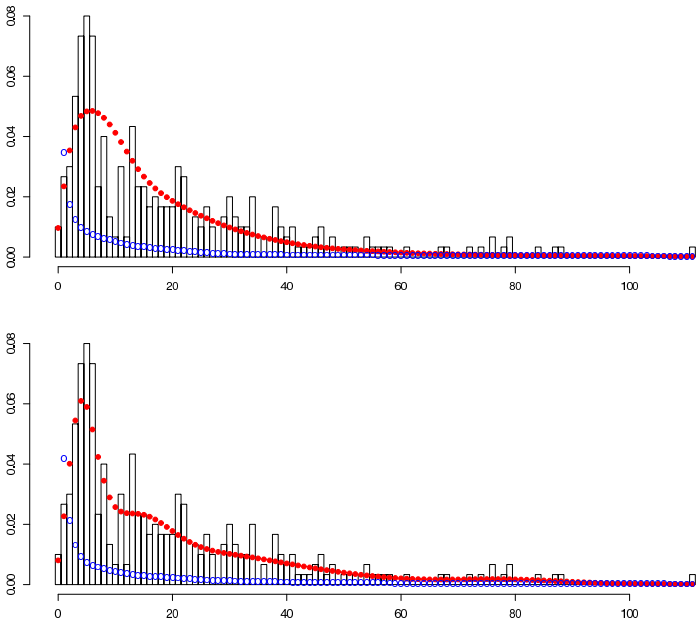


Figure 1: *Prior (blue) and posterior (red) predictive distribution from REP model (top) and BNP model (bottom).*

Antoniak (1974) derived an important result for the posterior distribution of the random mixing distribution $G$. It turns out that for $G \sim DP(\alpha, G_0(\psi))$, the posterior of $G$ is as follows:

$$(G|\text{data}) \sim \int P(G|\theta, \alpha, \psi)\mathrm{d}Q(\theta, \alpha, \psi \mid \text{ data }), \tag{5}$$

where $P(G \mid \theta, \alpha, \psi)$ is also a DP with parameters $\alpha^{'} = \alpha + n$ and

$$G_0^{'}(\cdot|\psi) = \frac{\alpha}{\alpha + n} G_0(\cdot|\psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{(-\infty, \theta_i]}(\cdot), \tag{6}$$

and $Q(\theta, \alpha, \psi \mid \text{ data })$ is the joint posterior distribution. Using (5), (6) along with the definition of DP we obtain posterior sample paths from $P(G \mid \text{data})$ in a computationally efficient way.

## 2 Parametric vs BNP models: the price of model uncertainty

Posterior estimates of the means of random effects distribution $G$ along with the 90% pointwise uncertainty bands are shown in Figure 2. It is obvious that the REP model can't capture the skewness and bimodality of the CDF (of the distribution of random effects, $G$), what is not surprising since REP assumes a Gaussian here. Yet, what is somewhat remarkable in a negative way is the very narrow uncertainty bands. On the other hand, the BNP model captures well both non-standard shapes of the CDF-s as expected, albeit with wider uncertainty bands around the mean estimate.

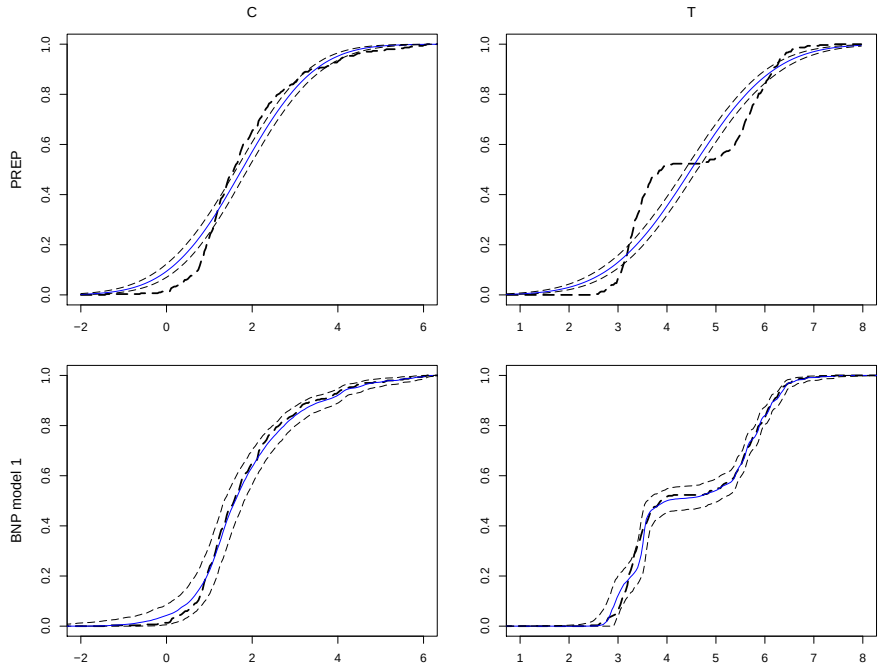


Figure 2: *Posterior MCMC estimates of the means of random effects distributions $G$, with* 90% *uncertainty bands. REP model, first row; BNP model, second row. Data sets generated using a model with skewed (left panels) and bimodal (right panels) distributions of random effects. The CDF-s of these (true) distributions are represented with thick dashed lines.*

We have seen that when REP is incorrect model, it continues to yield narrower uncertainty bands that fail to include the truth, whereas BNP model adapts successfully to the data-generating mechanism, as is illustrated in Figure 2. However, the Gaussian assumption on the latent variables scale in the REP model, although wrong, can make the model look plausible when it's not: Diagnostic checking of REP model would make it look appropriate when it's not; by contrast BNP correctly captures the bimodality or skewness of the random effects distribution.

One way to pay the right price for conducting a data-analytic search to arrive at a final parametric model is three-way cross-validation (3CV) which proceeds along the following lines: (1) Partition data at random into *three* subsets $S_i$, of size $n_i$ (respectively). (2) Fit tentative {likelihood + prior} to $S_1$. Expand initial model in feasible ways suggested by the data exploration using $S_1$. Iterate until fit is good (for example). (3) Use final model (fit to $S_1$) from (2) to create predictive distributions for all data points in $S_2$. Compare actual outcomes with these distributions, checking the predictive performance. Go back to (2), change likelihood or re-tune the prior as necessary, to get good calibration. Iterate until the predictive performance is OK (for example). (4) Announce final model (fit to $S_1 \cup S_2$) from (3), and report predictive calibration of this model on data points in $S_3$ as an indication of how well it would perform with new data.

In practice, with large $n$ we probably only need to do this once, whereas with small and moderate $n$ it may be necessary to repeat (1–4) several times and (perhaps) combine results in some way (for example, through model averaging). Note again that $n_3$ observations in $S_3$

Table 1

| | REP | | | | DPMP | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | Area | $r_A$ | MaxDiff | $r_D$ | Area | $r_A$ | MaxDiff | $r_D$ |
| 200 | 0.2256 | 1.403 | 0.11510 | 1.440 | 0.5556 | 1.160 | 0.3910 | 1.415 |
| 400 | 0.1608 | 1.433 | 0.07992 | 1.372 | 0.4788 | 1.141 | 0.2763 | 1.007 |
| 800 | 0.1122 | 1.427 | 0.05827 | 1.456 | 0.4195 | 1.090 | 0.2745 | 1.123 |
| 1600 | 0.0786 | | 0.04002 | | 0.3849 | | 0.2445 | |

are *not* to be used in summarizing inferential uncertainty about the quantities of interest but are instead used to estimate calibration of the data-analytic modeling process.

We need to find a way to determine or estimate sizes $n_i$ of three data subsets $(S_1, S_2, S_3)$. In order to approach this task of quantifying the price of model uncertainty it is useful (a) to regard Bayesian parametric models as just BNP models with a stronger prior. For example: REP model takes $G \equiv N(\mu, \sigma^2)$ while DP mixture model takes $G \sim DP(\alpha G_0), G_0 \equiv N(\mu, \sigma^2)$. Notice that larger sample sizes and stronger prior information often lead to narrower uncertainty bands.

Therefore, it is natural that a BNP model, on account of its vague prior on a large space of distribution functions, would require more data (sample size $n_{BNP}$) to achieve (about) the same inferential or predictive accuracy as the best-fitting (best-predictive) *parametric* model (in terms of sample sizes, $n_{BNP} >$ sample size $= n = n_{Param}$. It is then reasonable to recommend $n_3 = n(1 - n/n_{BNP})$ as the size of the calibration subset $S_3$. Combining this with the typical cross-validation practice that you should put about twice as much data in the modeling subset as in the validation subset yields

$$(n_1, n_2, n_3) = \left[\frac{2n^2}{3n_{BNP}}, \frac{n^2}{3n_{BNP}}, n\left(1 - \frac{n}{n_{BNP}}\right)\right]. \tag{7}$$

Therefore, for a data set with $n = 1000$ observations, if it takes about $n_{BNP} = 1200$ observations to achieve BNP accuracy equivalent to that of the best parametric model on the main quantities of interest, the subsets $S_i$ should have about $(550, 275, 175)$ observations in them.

Implementing this idea (obviously) requires estimating $n_{BNP}$. As a data-generating mechanism we use a REP model (with Gaussian $G$); and generate four samples of sizes $n = (200, 400, 800, 1600)$. To quantify the effect of (doubling) sample size, we compute (1) the areas between the 0.05 and 0.95 point-wise quantiles of the posterior realization of the CDF-s of $G$, and (2) the maximum differences between two quantiles. The results are summarized numerically in Tables 1 and 2. Figure 3 shows estimates of the 90% uncertainty bands of the posterior distribution of the CDF of $G$ for parametric and BNP model and different sample sizes.

We see that the REP model learns about $G$ at a substantially faster rate than the DPMP model. Noting the values of $r_A$ and $r_D$, the ratios of the consecutive values of"Area" and "MaxDiff" it appears that the REP learning rate follows a square root law, but the DPMP rate does not. However, if the data-generating mechanism was non-REP the REP model would continue to "learn" the wrong CDF at a same $\sqrt{n}$ rate, whereas the DPMP model would (somewhat slower) learn the right $G$.

Besides looking at the scale of latent variables, a similar comparison can be made on the

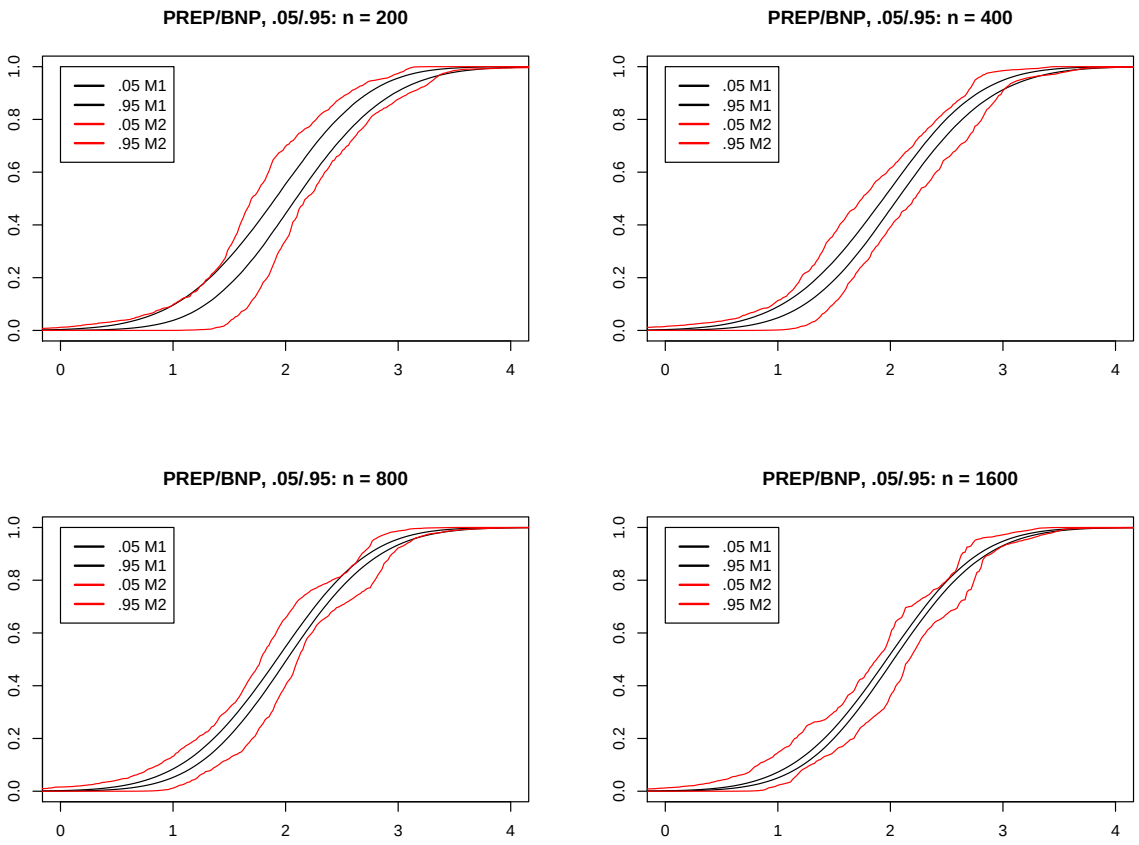


Figure 3: *95% point-wise uncertainty bands of posteriors of $G$, produced by the REP model (smooth & black) REP, and DPMP model (jagged & red).*

data scale and for that purpose we use the mean functional:

$$E(y \mid G) = \sum_{y=0}^{\infty} yF(y;G) = \sum_{y=0}^{\infty} y \int \text{Poisson}(y;\theta)G(\mathrm{d}\theta) = \int e^{\theta}G(\mathrm{d}\theta).$$

The mean functional has a closed form in case of REP model, whereas for the DPMP model we use MCMC draws from the joint posterior distribution of all parameters to compute it. The results for $E(y \mid G)$ are summarized in Table 2. It was unexpected to see that DPMP appears to learn about the posterior mean on the data scale at a faster rate than REP, although the difference between the two decreases for larger sample sizes. The result is counterintuitive, but an explanation may be given based on how the standard MCMC estimate of the posterior mean on the data scale is computed:

$$u_j = \sum_{k=1}^{K} \exp(t_k)\left[G_j(t_k) - G_j(t_k^-)\right] \tag{8}$$

for each MCMC iteration $j$, where $\{t_1, \ldots, t_K\}$ is a grid of points at which $G_j(\cdot)$, the current MCMC iteration estimate of $G$, is evaluated; the many flat segments in $G_j$ when the sample size is small can result in the uncertainty assessment on the low side.

Table 2

| | 90% Interval Width For $E(y|\text{data})$ | | | |
|---|---|---|---|---|
| $n$ | REP | | DPMP | |
| 200 | 1.793 | 1.433 | 1.679 | 1.424 |
| 400 | 1.251 | 1.564 | 1.179 | 1.483 |
| 800 | 0.800 | 1.396 | 0.795 | 1.438 |
| 1600 | 0.573 | | 0.553 | |

## Conclusion

To summarize the results, we can say that BNP models adapt well at latent and data levels and have superior predictive performance. It is interesting to see that a weaker prior information (provided by specifying priors on space of distributions in BNP) does not necessarily lead to weaker inferential statements on the data scale. Stronger prior information (when wrong, but difficult to diagnose) can lead to wrong inference in a somewhat striking manner.

Inferential uncertainty measured on the latent scale was smaller for parametric models and decreased with sample size, however, the uncertainty on the data scale was smaller for the BNP model. It means that this search for data equivalence between parametric and BNP models leads eventually in oposite directions and cannot be used to estimate the desired amount of data to use for calibration in DA approach.

The concept of data equivalence, if it worked, could lead to a fairly general way of quantifying the price of uncertainty for a data-driven search of the model space. However, the structure of the space of latent variables in BNP models changes non-trivially with the sample size and also reflects the features of the data set (such as skewness and multimodality), making the comparison with parametric problems a challenge. In general, $p(y \mid x) = \int p(y \mid x, M)p(M \mid x)dM$, where $M$ is a space of models, $p_1(M)$ may be a weaker prior than $p_2(M)$, and yet $p_1(M)$ may concentrate on models with better predictive accuracy, $p(y \mid x, M)$ than $p_2(M \mid x)$ does, leading to stronger inference from $p_1(y \mid x)$ than from $p_2(y \mid x)$.

## References


Antoniak, C. (1974). "Mixtures of Dirichlet processes with applications to nonparametric problems." *Annals of Statistics*, 2: 1152–1174.

Blackwell and MacQueen (1973). "Ferguson distributions via Plya urn schemes." *Annals of Statistics*, 1: 353–355.

Ferguson, T. (1974). "Prior distributions on spaces of probability measures." *Annals of Statistics*, 2: 615–629.

Krnjajić, M., Kottas, A., and Draper, D. (2008). "Parametric and nonparametric Bayesian model specification: a case study involving models for count data." *Computational Statistics and Data Analysis*, 52: 2110–2128.