20 Bayesian model specification: heuristics and examples

DAVID DRAPER

20.1 Introduction

You (a person wishing to reason sensibly in the face of uncertainty: [12]) are about to begin work on a new scientific problem $\mathbb{P}$. You begin by identifying θ, the unknown aspect of $\mathbb{P}$ of principal interest; in the story I wish to tell here, θ could be just about anything (e.g., a map precisely locating the highest and lowest points on the surface of a newly-discovered Earth-like extra-solar planet), but (for concreteness) think of $\theta = (\theta_1, \ldots, \theta_k)$ as a vector in $\Re^k$ (all finite-dimensional unknowns can be expressed in this way). You take stock of Your resources and realize that it's possible to obtain a new dataset D to decrease Your uncertainty about θ; again, D could be just about anything (e.g., a surveillance-camera video record of a crime, offering a partial identification of the perpetrator), but (again, for concreteness) think of $D = (y_1, \ldots, y_n)$ as a vector in $\Re^n$ (all datasets can be expressed in this way). Your other source of information relevant to solving $\mathbb{P}$ is a set $\mathcal{B}$ of (true/false) propositions, all regarded by You as true, describing the scientific context of $\mathbb{P}$ and the nature of the data-gathering process. (An example of a proposition in $\mathcal{B}$ from (e.g.) the field of history is as follows: $\{(y_1, \ldots, y_n)$ is a random sample of size n from the population $\mathcal{P}$ of all words in essay 19 of the *Federalist Papers*$\}$ [20], with θ as the unknown author of the essay (among Alexander Hamilton, James Madison, and John Jay).) At design time (i.e., when You're still contemplating how to obtain D), You notice that the existence of D at analysis time (i.e., after D has arrived) partitions the overall information about θ into {information internal to D} and {information external to D}, and this means that (at analysis time) You'll face a fundamental question: how should the information about θ both internal and external to D be combined, to create an optimal summary of Your total information (and therefore an accurate audit of Your uncertainty) about θ?

Here's a simple but real example, to fix ideas. In 1962 and 1963 [11], two employees of the US *National Bureau of Standards* (now called the *National Institute of Standards and Technology*) made $n = 100$ weighings of a block of metal called *NB10*—given this name because it was supposed to weigh 10 grams—under conditions that were as close as humanly possible to the statistical ideal of independent, identically distributed (IID) sampling from the population $\mathcal{P}_{\mathrm{NB10}} =$ {all possible weighings of *NB10* with the given apparatus}. Calling this problem $\mathbb{P}_{\mathrm{NB10}}$, here θ is evidently the 'true' weight of *NB10*, by which I mean the average of all the potential data values in $\mathcal{P}_{\mathrm{NB10}}$; D consists of the 100 weighings $y = (y_1, \ldots, y_n)$; and $\mathcal{B}$ contains the proposition $\{y$ is an IID sample from

$\mathcal{P}_{\text{NB10}}\}$ (along with background propositions known to be true from the context of $\mathbb{P}_{\text{NB10}}$, such as $\{\theta > 0\}$ and $\{\theta$ is close to 10 grams$\}$). In this problem the same fundamental question looms: how can an optimal summary of the total information about the weight of *NB10* be constructed?

The Bayesian approach to answering this *inferential* question, and to making *predictions* of future data D^* and *decisions* in the face of uncertainty, has been settled from a foundational perspective by de Finetti [2] and RT Cox [1]. Each of them proved a theorem, from different points of view about the meaning of probability: for de Finetti, probabilities arise from betting, and for Cox they're numerical expressions of information, in both cases about the truth status of propositions whose truth is unknown to You. The theorem says that if You specify two ingredients for inference and prediction—a probability distribution $p(D|\theta\, \mathcal{B})$ (usually referred to as Your *sampling distribution*) quantifying Your information about θ internal to D, and a probability distribution $p(\theta|\mathcal{B})$ (usually referred to as Your *prior distribution*) quantifying Your information about θ external to D—and two additional ingredients for decision-making—a set $\mathcal{A}$ of possible actions (usually referred to as Your *action space*) and a real-valued *utility function* $U(a, \theta^*)$ trading off the costs and benefits that will arise if You choose action a and θ takes the value θ^*—then (to obtain logically-internally-consistent inferences, predictions and decisions) You must combine the four ingredients according to the following three equations:

$$p(\theta|D\, \mathcal{B}) \propto p(\theta|\mathcal{B})\, p(D|\theta\, \mathcal{B}), \tag{20.1}$$

$$p(D^*|D\, \mathcal{B}) = \int_\Theta p(D^*|\theta\, D\, \mathcal{B})\, p(\theta|D\, \mathcal{B})\, d\theta, \tag{20.2}$$

$$a^* = \underset{a \in \mathcal{A}}{\text{argmax}}\, E_{(\theta|D\, \mathcal{B})}\, U(a, \theta) = \underset{a \in \mathcal{A}}{\text{argmax}} \int_\Theta U(a, \theta)\, p(\theta|D\, \mathcal{B})\, d\theta. \tag{20.3}$$

Here Θ is the set of possible values of θ; $p(\theta|D\, \mathcal{B})$ (usually referred to as Your *posterior distribution*) summarizes Your total information about θ and solves the inference problem; $p(D^*|D\, \mathcal{B})$, Your (posterior) *predictive distribution* for future data D^*, solves the prediction problem; and a^* solves the decision problem by *maximizing expected utility* (where the expectation is over Your posterior distribution $p(\theta|D\, \mathcal{B})$).

This is excellent, as far as it goes, but the original fundamental question has now been replaced by a new task that's almost as fundamental: how do You optimally specify the four ingredients {prior distribution, sampling distribution, action space, utility function} to be used in the three equations (20.1–20.3)? This task is *Bayesian model specification*, construed broadly. Sometimes this phrase is used more narrowly, to apply just to the sampling distribution, or just to {prior distribution, sampling distribution} if inference and/or prediction are the only goals. In the *NB10* problem, for instance, although there may be a subsequent decision with action space {replace *NB10* (because it doesn't actually weigh 10 grams), keep it}, I'll focus here on the inferential issue of the 'true' weight θ of *NB10*; in this problem let's call $M = \{p(\theta|\mathcal{B}), p(D|\theta\, \mathcal{B})\}$ Your *model* for (Your uncertainty about) θ.

To make the last paragraph meaningful I need to say what I mean by *optimal* Bayesian model specification, and this in turn depends on the following two-step argument:

- All Bayesian reasoning under uncertainty is based on $P(A|B) = \frac{P(A\, B)}{P(B)}$ for propositions A and B, and this is undefined if B is false; therefore
- **Rule 1:** You should try hard not to condition on propositions (a) that You know to be false and (b) that *may* be false.

This motivates the following terminology: in model specification, *optimal* = {to come as close as possible to the goal of [conditioning only on propositions rendered true by the context of the

problem and the design of the data-gathering process, while at the same time ensuring that the set of conditioning propositions includes all relevant problem context]}.

Achieving this goal seems hard; for example, a popular method of Bayesian model specification involves looking at the data to specify $p(D|\theta\ \mathcal{B})$—for example, with the *NB10* data You could make a normal quantile plot of the 100 observations and assume $\{(y_i|\theta\ \sigma^2\ \mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)\}$ for Your sampling distribution if the plot indicated approximate normality—but if You do this You'll be conditioning on a proposition that *seems* true on the basis of Your data analysis (see Rule 1(b)) but was not compelled by the problem context or data-collecting design. This approach can be regarded as a kind of 'cheating' in the model-specification process: You peek at the data to help guide this process away from conditioning on obviously false propositions, but the something-for-nothing bell in Your head is probably ringing—the very fact that You peeked may be an action that should come with a price-tag.

In this chapter I examine three methods that may be helpful in moving toward the optimal-model-specification goal described above: an approach called *Calibration Cross-Validation* (CCV) that helps You to pay the right price for the data-peeking mentioned in the previous paragraph (Section 20.2), and Bayesian non-parametric methods for specifying sampling distributions (Section 20.3) and prior distributions (Section 20.4).

20.2 Calibration cross-validation

Two paragraphs ago I mentioned that a common method for specifying the model $M = \{p(\theta|\mathcal{B}), p(D|\theta\ \mathcal{B})\}$ involves (a) looking at the data to identify an apparently reasonable choice for the sampling distribution $p(D|\theta\ \mathcal{B})$—call this particular choice S^*—and then (b) acting as if S^* is something that can safely be conditioned on in drawing inferences about θ. This clearly doesn't satisfy the definition of optimal model specification introduced in Section 20.1, because S^* didn't arise from the problem context or data-gathering design, and it's also likely to be deficient from a *calibration* point of view (in this chapter, a *well-calibrated* inferential process is one that, informally, gets the right answer about as often as it claims to do so): the S^* approach uses the full dataset twice (once to find S^*, and again to draw inferential conclusions about θ based on S^*). The mis-calibration consequences of the S^* approach will generally be that Your nominal $100(1-\gamma)\%$ inferential intervals for (univariate components of) θ and predictive intervals for (univariate components of) future datasets D^* will include the actual values less than $100(1-\gamma)\%$ of the time.

A natural approach to improving on the calibration performance of the S^* method for sampling-distribution specification is two-component *cross-validation* (CV), undertaken in three steps: first (1) You partition D exchangeably (see Section 20.3) into (mutually exclusive and exhaustive) modelling and validation subsets—call them $\mathbb{M}$ and $\mathbb{V}$, respectively; then (2) You explore a variety of models with the data in $\mathbb{M}$, eventually settling on one or more that appear to fit the data well; and then finally (3) You see how well the model(s) from (2) validate on the data in $\mathbb{V}$, for example by constructing $100(1-\gamma)\%$ predictive intervals (based on the data in $\mathbb{M}$) for all of the data values in $\mathbb{V}$ and seeing what percentage of these intervals contain the actual observations. (The S^* approach could be considered a kind of one-component CV, in which modelling and validation take place on the same data.)

Two-component CV (2CV) is clearly a big improvement on the S^* method, but what happens if the model(s) in step (2) don't validate well in step (3)? This occurs more often than You would like it to, and is an embarrassment for 2CV. The natural thing to do is to go back to step (2), re-modelling and re-validating in step (3), iterating (2) and (3) until You finally *do* have one or more models that validate well in $\mathbb{V}$, but You now notice that You've painted Yourself into a corner: You don't have

any pristine data values left to see how well the iterative modelling *process* calibrates on data not used in that process. This motivates *calibration cross-validation* (CCV; [4]): going out one more term in the Taylor series, so to speak, the idea is to

(a) partition the data into modelling ($\mathbb{M}$), validation ($\mathbb{V}$) and calibration ($\mathbb{C}$) subsets;
(b) use $\mathbb{M}$ to explore a variety of models until You've found one or more plausible candidates, which You can collect in an *ensemble* $\mathcal{M} = \{M_1, \ldots, M_m\}$;
(c) see how well the models in $\mathcal{M}$ predictively validate in $\mathbb{V}$;
(d) if none of them do, iterate (b) and (c) until You do get good validation, and
(e) fit the best model in $\mathcal{M}$ (or, better, use *Bayesian model averaging* (see, e.g., [19] and [3]) with the entire ensemble $\mathcal{M}$) on the data in ($\mathbb{M} \cup \mathbb{V}$), and report both (i) inferential conclusions based on this fit and (ii) the quality of predictive calibration of Your model/ensemble on the data in $\mathbb{C}$.

The goal with this method is both

(1) a good answer, to the main scientific question, that has paid a reasonable price for *model uncertainty* (the inferential answer is based only on ($\mathbb{M} \cup \mathbb{V}$), not the entire dataset, making Your uncertainty bands wider than those from an S^* analysis), and
(2) an indication of how well calibrated {the iterative fitting process yielding the answer in (1)} is in the calibration subset $\mathbb{C}$, which is a good proxy for future data.

You can use Bayesian decision theory [4] to decide how much data to put in each of $\mathbb{M}$, $\mathbb{V}$ and $\mathbb{C}$: the more important calibration is to You, the more data You want to put in $\mathbb{C}$, but only up to a point, because getting a good answer to the scientific question is also important. I've found that (0.5, 0.25, 0.25) is often a reasonable allocation of data fractions into ($\mathbb{M}$, $\mathbb{V}$, $\mathbb{C}$), and that's what I'll use here. In the rest of this subsection I illustrate the use of CCV on the *NB10* dataset, which is summarized in the top part of Table 20.1 (values are expressed in micrograms below the nominal weight of 10 g).

I randomly partitioned the 100 *NB10* data values into the ($\mathbb{M}$, $\mathbb{V}$, $\mathbb{C}$) subsets of sizes (50, 25, 25) given in the bottom part of Table 20.1 (for greatest stability of conclusions, this random partitioning should be repeated a number of times, with CCV performed in parallel on the repetitions and the results combined appropriately (see [4] for details); in the interests of brevity, here I only show results with the partition in Table 20.1). Step (b) of CCV now involves exploratory modelling with the data in $\mathbb{M}$.

Given the *NB10* problem context, it's natural to begin by fitting the parametric Gaussian model

$$M_1: \left\{ \begin{array}{lcl} (\theta\, \sigma^2 | \mathcal{B}) & \sim & p(\theta\, \sigma^2 | \mathcal{B}) \\ (y_i | \theta\, \sigma^2\, \mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \tag{20.4}$$

for $i = 1, \ldots, n$. At the point at which these 100 weighings of *NB10* were performed in 1962–63, it's likely that workers at the National Bureau of Standards (NBS) already knew quite a bit about the weight of this block of metal, but here I'm going to illustrate the analysis from the viewpoint of someone (like me, and probably You) who has little information external to the present dataset D about the actual weight of *NB10* or the accuracy of the NBS weighing process. To make this state of information operational, I used the diffuse prior $p(\theta\, \sigma^2 | \mathcal{B}) = p(\theta | \mathcal{B})\, p(\sigma^2 | \mathcal{B})$, with $(\theta | \mathcal{B}) \sim N(0, 10^6)$ and $(\sigma^2 | \mathcal{B}) \sim \Gamma^{-1}(0.001, 0.001)$ (other diffuse prior specifications yielded nearly identical conclusions). With this prior, under the Gaussian model (20.4), (i) the marginal

Table 20.1 Top: A raw frequency distribution of n = 100 weighings of *NB10*; bottom: the random CCV partition illustrated here (with the data values in each component sorted).

Value	375	392	393	397	398	399	400	401
Frequency	1	1	1	1	2	7	4	12
Value	402	403	404	405	406	407	408	409
Frequency	8	6	9	5	12	8	5	5
Value	410	411	412	413	415	418	423	437
Frequency	4	1	3	1	1	1	1	1

$\mathbb{M}$:	375	399	399	399	399	400	400	400	401	401	401	401	401
	401	402	402	402	402	402	402	403	403	403	403	403	404
	404	404	404	404	404	404	405	405	405	406	406	406	406
	406	407	407	407	408	408	408	409	410	411	437		
$\mathbb{V}$:	393	397	398	399	400	401	401	402	403	404	405	406	406
	406	407	407	407	408	408	409	409	412	412	418	423	
$\mathbb{C}$:	392	398	399	399	401	401	401	401	402	404	405	406	406
	406	406	407	407	409	409	410	410	410	412	413	415	

posterior for θ is approximately Gaussian with mean 403.8 and standard deviation (SD) 1.00, (ii) a 95% central posterior interval for θ runs from 401.8 to 405.8, (iii) the marginal posterior for σ has a moderately long right-hand tail (as You would expect for a scale parameter) with mean 7.06 and SD 0.730, (iv) the posterior predictive distribution for a future observation is approximately Gaussian with mean 403.8 and SD 7.17, and (v) the 95% central posterior predictive interval for the next data point runs from 389.7 to 417.9.

It's now interesting to see how well calibrated the Gaussian model is on the data set used to fit it. The left panel of Figure 20.1 presents a *calibration plot* based on the data in $\mathbb{M}$, comparing nominal and actual coverage of $100(1-\gamma)\%$ predictive intervals for $\gamma = (0.01, 0.02, \ldots, 0.99)$. You can see that the Gaussian model produces predictive intervals that are sharply conservative; for example, at all nominal levels from 70% to 95%, the actual coverage is 96%. The right panel of Figure 20.1 gives a normal quantile plot of the data in $\mathbb{M}$, which identifies the reason for the poor validation of the Gaussian model: the distribution is unimodal and close to symmetric but has substantially heavier tails than the Gaussian, and this has led in the Gaussian framework to a large estimate of σ. By way of a second, improved model this suggests a t sampling distribution, as in

$$M_2: \left\{ \begin{array}{lcl} (\theta\, \sigma^2\, \nu|\mathcal{B}) & \sim & p(\theta\, \sigma^2\, \nu|\mathcal{B}) \\ (y_i|\theta\, \sigma^2\, \nu\, \mathcal{B}) & \overset{\text{IID}}{\sim} & t_\nu(\theta, \sigma^2) \end{array} \right\}. \tag{20.5}$$

This model is easy to fit via MCMC with slice sampling; 100 000 monitoring iterations took 15 seconds at 3.3 GHz (this monitoring sample size produced Monte Carlo standard errors for all posterior summaries less than 0.01). Here I used the diffuse prior $p(\theta\, \sigma^2\, \nu|\mathcal{B}) = p(\theta|\mathcal{B})\, p(\sigma^2|\mathcal{B})\, p(\nu|\mathcal{B})$, with the same marginal priors as in the Gaussian model for θ and σ^2 and with $(\nu|\mathcal{B}) \sim$ Uniform(1.0, 10.0) (the right endpoint was chosen to be large enough to avoid truncation of the likelihood; other values that avoid truncation give similar results).

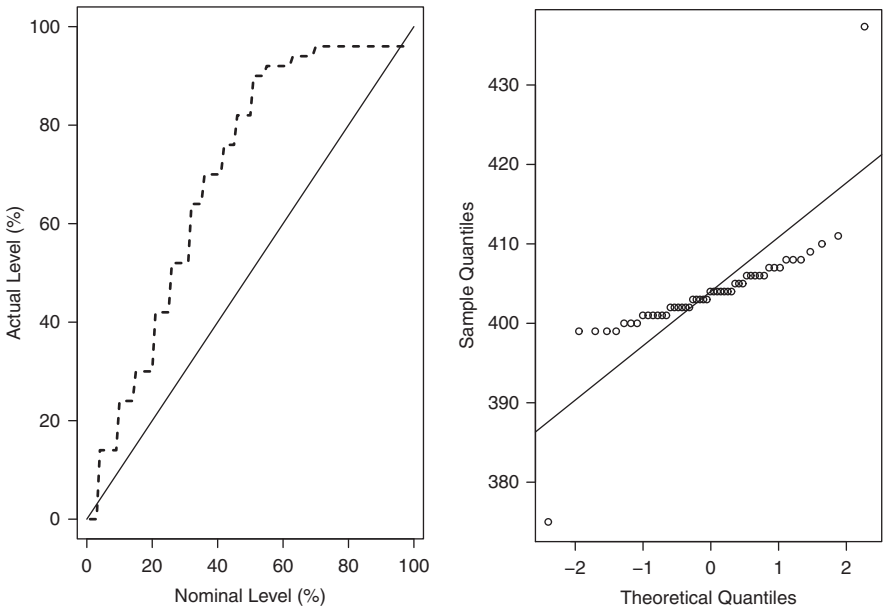

Figure 20.1 Left panel: calibration curve (dotted line) for the Gaussian model (20.4) fit to the *NB10* data in $\mathbb{M}$ and validated in $\mathbb{M}$ (the solid line represents the target behaviour under good calibration); right panel: normal quantile plot of the data in $\mathbb{M}$.

The results from fitting the t model (20.5) to $\mathbb{M}$ are as follows: (i) the marginal posterior for θ is again approximately Gaussian, but this time with mean 403.4 and SD 0.50; (ii) a 95% central posterior interval for θ runs from 402.5 to 404.4; (iii) the marginal posterior for σ again has moderate positive skew, this time with mean 2.73 and SD 0.46; (iv) the marginal posterior for ν has a substantial right-hand tail, with (mode, median, mean) = (2.31, 2.44, 2.60) and SD 0.91; (v) the posterior predictive distribution for a future observation is approximately Gaussian with mean 403.4 and SD 2.80; and (v) the 95% central posterior predictive interval for the next data point runs from 397.9 to 409.0. Note how much smaller both the inferential uncertainty about θ and the predictive uncertainty about future observations are with the t model than with the Gaussian sampling distribution; this is a consequence of the Gaussian having minimal Fisher information for location among all symmetric unimodal sampling distributions on $\Re$. The much smaller values for σ are because observations from a $t_\nu(\theta, \sigma^2)$ sampling distribution have variance $\frac{\nu}{\nu-2}\sigma^2$ (i.e., scale and shape are confounded in this model).

Is the t model better than the Gaussian for the data in $\mathbb{M}$? There are a number of ways to answer this question; the one I like best [5] involves *full-sample log scores*. The idea, with a univariate dataset $D = y = (y_1, \ldots, y_n)$ (such as the *NB10* weighings) and models M_j (here $j = 1, 2$) to be compared, involves computing

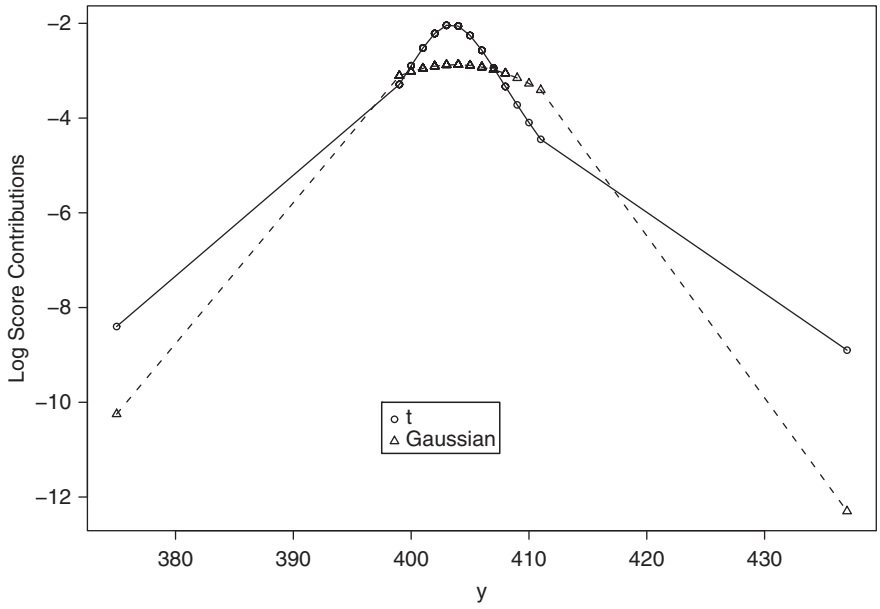

Figure 20.2 Contributions to the overall LS_{FS} values for each model from each observation; triangles (solid curve) and circles (dotted curve) track the Gaussian and t models, respectively.

$$LS_{FS}(M_j|D\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|D\,M_j\,\mathcal{B}) \tag{20.6}$$

and favouring the model with the bigger log score LS_{FS}. Computation of LS_{FS} is straightforward; when parametric model M_j with parameter vector η_j is fit via MCMC, the predictive ordinate $p(y^*|D\,M_j\,\mathcal{B})$ in LS_{FS} can be approximated as follows. With m identically distributed (not necessarily independent) MCMC monitoring draws $(\eta_j)_k^*$ from $p(\eta_j|D\,M_j\,\mathcal{B})$,

$$\begin{aligned} p(y^*|D\,M_j\,\mathcal{B}) &= \int p(y^*|\eta_j\,M_j\,\mathcal{B})\,p(\eta_j|D\,M_j\,\mathcal{B})\,d\eta_j \\ &= E_{(\eta_j|D\,M_j\,\mathcal{B})}\,p(y^*|\eta_j\,M_j\mathcal{B}) \\ &\doteq \frac{1}{m}\sum_{k=1}^{m} p[y^*|(\eta_j)_k^*\,M_j\,\mathcal{B}]. \end{aligned} \tag{20.7}$$

Applying this method to models M_1 ((20.4), Gaussian) and M_2 ((20.5), t) with the data in $\mathbb{M}$ yields LS_{FS} values of -3.30 and -2.86, respectively; this represents a (sharp) preference for the t model. Figure 20.2 shows the individual contributions, from each data value in $\mathbb{M}$, to the overall LS_{FS} values from the Gaussian and t models. It's evident that the t model fits better both in the tails (where the most influential observations are from the Gaussian point of view) and in the centre (where most of the data values are); in fact, 80% of the data values in $\mathbb{M}$ are predicted better by the t model than by the Gaussian.

Next question: is the t model good enough to stop looking for a better model? The answer is complicated here by the small sample sizes in each of the $\mathbb{M}$ and $\mathbb{V}$ partition components. Figure 20.3

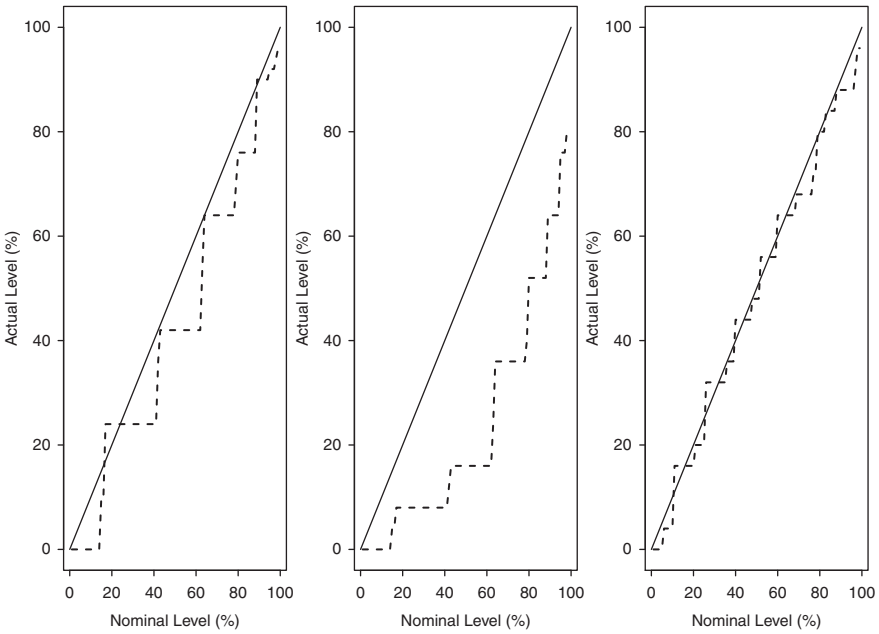

Figure 20.3 Calibration plots for the *t* model (20.5), fit to the data in $\mathbb{M}$ and validated in $\mathbb{M}$ (left panel); fit to $\mathbb{M}$ and validated in $\mathbb{V}$ (centre); and fit to $\mathbb{V}$ and validated in $\mathbb{V}$ (right).

gives calibration plots for the *t* model with three different data configurations: fit to the data in $\mathbb{M}$ and validated in $\mathbb{M}$ (left panel); fit to $\mathbb{M}$ and validated in $\mathbb{V}$ (centre; this corresponds to step (c) in the CCV algorithm); and fit to $\mathbb{V}$ and validated in $\mathbb{V}$ (right). *Internal* validation (evaluating the fit on the same dataset used to create the fit, as in {fit to $\mathbb{M}$, validate in $\mathbb{M}$}, which could be abbreviated $\mathbb{M} \rightarrow \mathbb{M}$, and similarly $\mathbb{V} \rightarrow \mathbb{V}$) ranges from barely adequate (the left panel) to excellent (the right display), but *external* validation (evaluating the fit on new data not used in the fitting process, as in $\mathbb{M} \rightarrow \mathbb{V}$, in the centre) is abysmal, with the opposite calibration problem (predictive intervals that are not wide enough) from that exhibited by the Gaussian $\mathbb{M} \rightarrow \mathbb{M}$ model in the left panel of Figure 20.1. With only 50 observations in $\mathbb{M}$ and 25 in $\mathbb{V}$, the parameter estimates from the *t* model are quite different when it's applied separately to $\mathbb{M}$ and $\mathbb{V}$ (see the first two rows in Table 20.2 below). Another way to put the difficulty, looking at the full *NB10* dataset in Table 20.1, is that there are 'only' 3 outliers in the entire dataset (namely, the observations 375, 423 and 437), and the presence or absence of any one of these outliers in $\mathbb{M}$ or $\mathbb{V}$ is, by its very nature, highly influential for the parameter estimates. As mentioned previously, [4] performs the obvious analysis to remedy this problem—repeat the CCV algorithm across many random ($\mathbb{M}$, $\mathbb{V}$, $\mathbb{C}$) partitions and average the results—which, in the interests of brevity, I do not reproduce here; the conclusion from this broader analysis is that the *t* model is a good basis for stopping the iterative step (d) in CCV and proceeding to the final step (e).

The third row in Table 20.2 gives posterior summaries from fitting the *t* model to the 75 observations in ($\mathbb{M} \cup \mathbb{V}$), and the left panel in Figure 20.4 gives the calibration plot for this fit when applied to $\mathbb{C}$. You can see that the model's validation is not perfect, again in part because of the small sample sizes: for instance, the nominal 95% predictive interval runs from 397.1 to 410.9 and includes only 88% of the data values in $\mathbb{C}$. Although it's not part of the CCV algorithm to do so, for comparison purposes the final row of Table 20.2 summarizes the fit of the *t* model to the entire

Table 20.2 Parameter and predictive summaries from fitting the t model separately to the $\mathbb{M}$ and $\mathbb{V}$ partition components, to the merged dataset ($\mathbb{M} \cup \mathbb{V}$), and to the entire dataset D; y^* is a future data value.

Data partition	Sample size	Posterior mean (SD)			
		θ	σ	ν	y^*
$\mathbb{M}$	50	403.4 (0.50)	2.73 (0.46)	2.60 (0.91)	403.4 (2.80)
$\mathbb{V}$	25	405.3 (1.23)	5.31 (1.12)	5.73 (2.39)	405.3 (5.54)
$(\mathbb{M} \cup \mathbb{V})$	75	404.0 (0.50)	3.42 (0.47)	2.88 (0.95)	404.0 (3.48)
D	100	404.3 (0.47)	3.85 (0.45)	3.56 (1.18)	404.3 (3.91)

dataset D, and the right panel in Figure 20.4 displays the calibration plot that results when the t model is fit to, and validated in, all of D; this shows what someone using the S^* approach would conclude, both about the parameters and about the quality of the model fit. The right panel of Figure 20.4 provides a somewhat rosier view of the quality of the t model than the left panel, and is therefore somewhat misleading about the calibration performance of the iterative modelling process leading to the results of the S^* method.

On the basis of the CCV approach to dealing with specification uncertainty about the sampling distribution in the *NB10* problem, I would draw the following conclusions:

(A) The block of metal called *NB10* weighed (in 1963) about 404.0 micrograms below the nominal weight of 10 grams, give or take about 0.50 micrograms, and a 95% interval for its weight runs from 403.0 to 404.9; and

(B) the iterative modelling process leading to the inferential conclusion in (A) is somewhat over-confident in its ability to predict future data values not used in the model-fitting, with nominal 95% predictive intervals for future observations including the actual data values about 88% of the time, give or take about $100\sqrt{\frac{(0.88)(0.12)}{25}}\% \doteq 6.5\%$.

20.3 Bayesian nonparametric sampling-distribution specification

In the *NB10* problem, at design time (before any data have been collected), and with no covariate information that would serve to distinguish one observation from another, upon reflection (following de Finetti [2]) You would notice that Your uncertainty about the *NB10* weighings $D = (y_1, \ldots, y_n)$ is *exchangeable*, in the usual sense that Your predictive distribution $p(y_1 \ldots y_n|\mathcal{B})$ is invariant under permutation of the order in which the data values are observed. Moreover, if the weighing process were to be continued indefinitely (still with no covariate information), yielding the entire population $\mathcal{P}_{\text{NB10}} = (y_1, y_2, \ldots)$, Your predictive distribution $p(y_1\, y_2 \ldots|\mathcal{B})$ would still be exchangeable (in the sense that exchangeability would hold for any finite subset of $\mathcal{P}_{\text{NB10}}$). In settings such as this, de Finetti [2] proved a celebrated theorem that says (slightly informally)

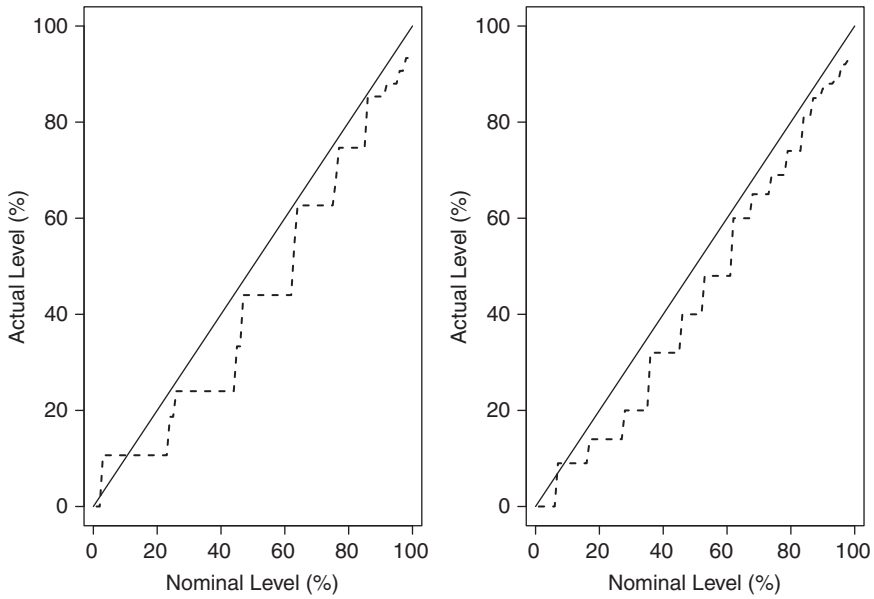

Figure 20.4 Calibration plots for the t model (20.5), fit to the data in $(\mathbb{M} \cup \mathbb{V})$ and validated in $\mathbb{C}$ (left panel); and fit to, and validated in, the entire data set (right panel).

that all logically-internally-consistent predictive distributions $p(y_1 \ldots y_n|\mathcal{B})$ are expressible hierarchically as

$$\left\{ \begin{array}{rcl} (G|\mathcal{B}) & \sim & p(G|\mathcal{B}) \\ (y_i|G\,\mathcal{B}) & \overset{\text{IID}}{\sim} & G \end{array} \right\}, \tag{20.8}$$

where G is the empirical CDF of $(y_1, y_2, \ldots)$; here $p(G|\mathcal{B})$ is a prior distribution on the set $\mathcal{G}$ of all CDFs on $\Re$. This theorem founded the sub-field of *Bayesian non-parametric (BNP) modelling*, which concerns inference on functions such as G (and also—not addressed in this chapter—functions such as regression surfaces). At the time he proved the theorem, de Finetti didn't know how to put a scientifically meaningful prior on $\mathcal{G}$, but progress toward this goal—started 40–50 years ago by Freedman [10] and Ferguson [9]—culminated, in the work of people such as Escobar and West [7] and Lavine [18], with MCMC-based approaches to extracting information from the posterior distribution $p(G|D\,\mathcal{B})$, and BNP modelling has become increasingly routine in the past 15 years. This approach offers the possibility of optimal model specification (in the definitional sense in Section 20.1) in the *NB10* problem, because the judgement of exchangeability leading to model (20.8) arises directly from problem context; the only remaining issue with this approach is how to specify $p(G|\mathcal{B})$ in a manner that is both (a) accurately driven by the nature of the data-gathering process and (b) well-calibrated.

Two approaches to specifying $p(G|\mathcal{B})$ have by now been developed to the point that they're both scientifically useful and computationally tractable: *Dirichlet-process (DP) mixture modelling* [7, 9] and *Pólya-tree (PT) mixture modelling* (e.g., [13]). I'll concentrate here on Pólya trees; see, e.g., [17] for practical examples of DP modelling with count data. For a univariate sample $D = y = (y_1, \ldots, y_n)$ such as the *NB10* dataset, a natural PT mixture model would take the following form:

$$\left\{ \begin{array}{rcl} (y_i|G\mathcal{B}) & \overset{\text{IID}}{\sim} & G \qquad (i = 1, \ldots, n) \\ (G|\alpha\theta\sigma^2\mathcal{B}) & \sim & PT\left[\Pi_{N(\theta,\sigma^2)}, \mathcal{A}_\alpha\right] \\ (\alpha\theta\sigma^2|\mathcal{B}) & \sim & p(\alpha\theta\sigma^2|\mathcal{B}) \end{array} \right\}, \tag{20.9}$$

for an appropriately chosen prior distribution $p(\alpha\ \theta\ \sigma^2|\mathcal{B})$ on $(\alpha, \theta, \sigma^2)$.

The meaning of the expression $PT\left[\Pi_{G_0(\eta)}, \mathcal{A}_\alpha\right]$ is as follows. Rather generally in Bayesian work, prior distributions are specified through two main ingredients: a *prior estimate* of the thing receiving the prior distribution, and a *prior sample size* indicating how tightly concentrated the prior should be around the prior estimate. PT priors for a CDF G follow this pattern: $G_0(\eta)$ is the prior estimate or *centring distribution*, which will typically be a parametric family indexed (in this case) by the parameter vector η, and α acts like a prior sample size, in the sense that bigger (smaller) values of α lead to posterior distributions on G that are closer to (farther away from) the centring distribution. In model (20.9), $G_0(\eta)$ is the $N(\theta, \sigma^2)$ distribution; this is natural in the *NB10* problem for the same reason that the Gaussian sampling distribution appeared in model M_1 in Section 20.2.

This approach is referred to as PT mixture modelling because a point-mass prior on $(\alpha, \theta, \sigma^2)$ of the form $(\alpha = \alpha_0, \theta = \theta_0, \sigma^2 = \sigma_0^2)$ would correspond to fitting a single Pólya-tree prior for G, whereas a more realistic treatment of prior uncertainty about $(\alpha, \theta, \sigma^2)$—in which non-point-mass distributions are given to one or more elements of the $(\alpha, \theta, \sigma^2)$ vector—amounts to mixing over individual Pólya trees. For univariate outcomes, PT priors are based on binary partitions of $\Re$ with 2^m partition sets at level m of the tree, and act like random histograms; to get PT priors to directly model continuous data, strictly speaking the number of histogram bars has to become countably infinite, but in practice finite Pólya trees (with 2^M bars, for finite M, at the bottom level) are all that's needed, because the real-world process of measuring conceptually continuous outcomes always discretizes them anyway.

These days it's relatively straightforward to fit model (20.9) via MCMC with a Metropolis-within-Gibbs approach: the full-conditional distribution $p(G|D\ \alpha\ \theta\ \sigma^2\ \mathcal{B})$ turns out to be another Pólya tree, and then You can Metropolis-sample the other full-conditionals (such as $p(\theta|D\ G\ \alpha\ \sigma^2\ \mathcal{B})$). The ensemble of `R` functions called `DPpackage` [14], available from `CRAN`, contains several functions that can fit model (20.9), including `PTlm` and `PTdensity`, and `WinBUGS` code for this model is available from Tim Hanson; this permits attention to shift away from the MCMC details and toward the modelling, where several surprises await (in relation to Your experience with parametric modelling).

It's possible to put a prior distribution on α, but—with an eye on calibration, as in Section 20.2—You can instead regard α as a kind of tuning constant that You can vary across a range of fixed values to achieve good out-of-sample calibration. In the *NB10* problem, I again use a diffuse prior on (θ, σ^2)—Gaussian with huge variance for θ, $\Gamma^{-1}(\epsilon, \epsilon)$ for σ^2 with small positive ϵ—to quantify the information base of someone who knows little, external to the *NB10* dataset, about the weight of *NB10* or the accuracy of the weighing process.

As a first example of the results from the BNP approach to dealing with specification uncertainty about sampling distributions, I used `PTlm` on the entire *NB10* dataset with $\alpha = 1$ and $M = 6$, employing a burn-in of 5000 iterations (from starting values for μ and σ that were not far from their likely posterior means) and a monitoring run of 10 000 saved values after thinning by a factor of 20. (In `PTlm`, by default θ is identified as the *median* of the population empirical CDF.) The resulting 205 000 iterations took 4.5 minutes at 3.3 GHz, and initially yielded poor acceptance rates for the Metropolis steps for θ and σ^2. Iterative tuning of the proposal distribution SDs eventually yielded near-optimal univariate acceptance rates of 44–49%, at which point I examined the Monte-Carlo accuracy achieved by this MCMC sampling strategy. The saved iterations for θ behaved like draws

from an $AR_1(\rho_1)$ time series with a first-order autocorrelation $\hat{\rho}_1$ of +0.75, even after 20-fold thinning. From the usual expression

$$\widehat{MCSE}\left(\bar{\theta}^*\right) = \frac{\hat{\sigma}_\theta}{\sqrt{n^*}}\sqrt{\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}} \tag{20.10}$$

for the Monte Carlo standard error (MCSE) of the MCMC estimate $\bar{\theta}^* = \frac{1}{n^*}\sum_{j=1}^{n^*}\theta_j^*$ of the posterior mean of θ, where $\hat{\sigma}_\theta$ is the estimated posterior SD of θ and n^* is the number of saved monitoring iterations, it became clear that—with only 200 000 iterations going into the monitoring process—the MCSEs of the posterior mean and SD estimates were on the order of 0.08, which was too big for getting a good idea of the posterior SD of θ. To drive the MCSEs down to about 0.01, a monitoring run of 12 000 000 iterations (thinning by a factor of 200) was needed; this took about 3.9 hours at 3.3 GHz. The first surprise with BNP modelling is how much longer in clock time it can take to get results with decent Monte-Carlo accuracy, in relation to Your parametric-modelling experience; on reflection, this is perhaps not actually so surprising, for two reasons: (i) You're treating G as a nuisance parameter that has to be learned along with the main parameter(s) of interest (with $M = 6$ in Pólya trees, this is like learning an additional $2^6 = 64$ parameters (albeit rather highly correlated, so that the effective dimensionality of the learning process for G is probably on the order of a few dozen additional parameters)), and (ii) uncertainty about G is bound to create poorer mixing for θ and the other parameters You care about.

Figure 20.5 displays the marginal posterior distributions for θ (left panel) and σ (right panel) from this fitting of model (20.9), using default window-widths for the kernel density estimation. The second surprise with BNP modelling, when compared with parametric-modelling intuition, is how rough these posterior distributions are; on reflection this is once again perhaps not so

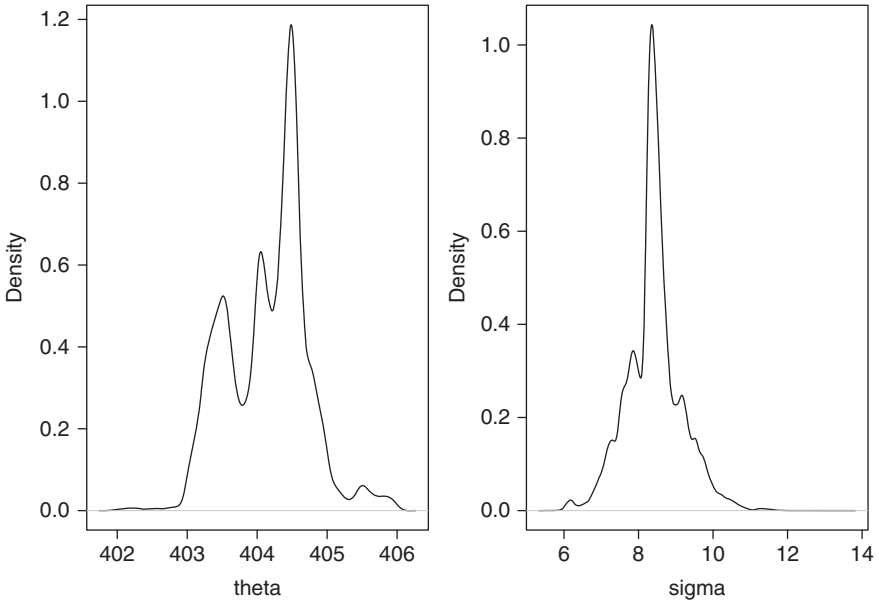

Figure 20.5 Marginal posterior distributions for θ (left panel) and σ (right panel), from fitting the Pólya-tree model (20.9).

startling (with the *NB10* data, the empirical CDF is itself quite rough, from the granularity of the observations). The marginal posterior for θ has a mean of 404.2 and an SD of 0.58, and the 95% central posterior interval runs from 403.1 to 405.4; these results are in reasonable agreement with those from the CCV approach in Section 20.2, with perhaps a bit more uncertainty about θ arising from what may be a better attempt to fully quantify uncertainty about G.

As a second example of the fitting of BNP models, to get a closer look at their calibration properties, I created an artificial dataset that had the same mean $\bar{y}$ and SD s as the *NB10* data but was (in a certain sense) as close to Gaussian as possible: observation i in this artificial dataset had the value $y_i' = \bar{y} + s\,\Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$, where Φ is the standard normal CDF. I then fit model (20.9) to this artificial dataset, with 12 different priors on $(\alpha, \theta, \sigma^2)$; all these priors were of the form $p(\alpha, \theta, \sigma^2|\mathcal{B}) = p(\alpha|\mathcal{B})\,p(\theta|\mathcal{B})\,p(\sigma^2|\mathcal{B})$, with $(\theta|\mathcal{B}) \sim N(0, 10^6)$ and $(\sigma|\mathcal{B}) \sim U(0, 20)$ in each case (the upper limit on the uniform prior on σ was again chosen to avoid likelihood truncation). My goal in this work was (a) to represent (as in all of the results in this section) the information base, external to the *NB10* data, of someone who knows little about the weight of *NB10* or the accuracy of the NBS measuring process, and (b) to examine the resulting posterior inferences about (θ, σ) as a function of various priors on α, which is also (to put it mildly) not a quantity strongly pinned down by information external to the *NB10* dataset. I used Tim Hanson's `WinBUGS` code for these runs; in this code θ is identified as the *mean* of the population empirical CDF.

Consider a point-mass prior on α that sets $\alpha = \alpha^*$ (say). At each iteration of the Metropolis-within-Gibbs sampling used to fit model (20.9), at the point at which θ^* and σ^* values have been drawn from $p(\theta\,\sigma|G\,D\,\mathcal{B})$, imagine standardizing the artificial data values y_i' to create $y_i'' = \frac{y_i' - \theta^*}{\sigma^*}$, and let G'' be the empirical CDF of the resulting y_i'' values. To complete the current scan of the sampler, the final step is to draw a CDF G^* from the Pólya-tree distribution $PT\left[\Pi_{G^\dagger}, \mathcal{A}_{\alpha^*}\right]$, where $G^\dagger$ is a weighted average of the standard normal CDF Φ and G'' with weights given by α^* and n (respectively). Thus as α^* grows (with n fixed at 100), with the artificial dataset examined here (in which the empirical CDF is as close to Gaussian as possible), You would expect the Pólya-tree results to more and more closely resemble those from fitting the parametric Gaussian model (20.4), with the same diffuse prior on (θ, σ) as above; the question is how quickly (as α^* increases) this convergence will occur.

Table 20.3 presents the results of these calculations. By way of priors on α I used a popular choice in BNP modelling—a variety of $\Gamma(a, b)$ priors on α (almost all of which had $b = 1$)—and I compared these with point-mass priors having the same prior means as the Gamma distributions; the bottom row of the table gives the parametric Gaussian results for further comparison. (All of the Monte Carlo standard errors for the values in this table were 0.01 or smaller.) You can see that the expected convergence has indeed occurred, but the interesting thing (and this is a third surprise from BNP modelling) is how large α needs to be to get (calibrationally correct) results that are close to those from the parametric model. With small α, even with $n = 100$ observations, with diffuse priors on θ and σ, the uncertainty about those two parameters imposed upon the BNP modelling, above and beyond the uncertainty about G, makes the BNP inferences extremely conservative. (Of course, to really pin this down You would have to create a simulation environment in which many Gaussian datasets were generated at random, rather than simply using the one "super-Gaussian" artificial dataset I used here; I intend to report on results from this broader simulation experiment elsewhere.)

The reason for the inferential conservatism in Table 20.3 with small α is that, in the BNP formulation, (θ, σ) and G are correlated in the posterior (especially when α is small), and uncertainty about G is therefore propagated into uncertainty about the parameters. As an example of these correlations, I monitored the posterior for G on an equally spaced grid of 200 points in the range $(\bar{y} \pm 3.5s)$, obtaining a vector $(G_1^*, \ldots, G_{200}^*)$ on each MCMC scan; with $\alpha = 1$, correlations

Table 20.3 Posterior summaries from fitting the Pólya-tree model (20.9) with an artificial Gaussian dataset having the same mean and SD as the *NB10* data, using a variety of prior distributions on α. In the first column, an integer k signifies $\alpha = k$, and $\Gamma_{a,b}$ is the $\Gamma(a, b)$ distribution. The last row gives results from fitting the parametric Gaussian model (20.4) to the same dataset, for comparison.

	Posterior summaries for					
	θ		σ		α	
α	Mean	SD	Mean	SD	Mean	SD
$\Gamma_{1,1}$	404.6	1.69	6.75	0.72	3.26	1.53
1	404.5	3.13	7.11	1.11	—	—
$\Gamma_{5,1}$	404.6	1.32	6.66	0.59	6.37	2.54
$\Gamma_{10,2}$	404.6	1.63	6.66	0.60	5.74	1.63
$\Gamma_{10,1}$	404.6	1.09	6.61	0.55	10.9	3.21
10	404.6	1.11	6.62	0.55	—	—
$\Gamma_{20,1}$	404.6	0.96	6.59	0.51	20.6	4.53
$\Gamma_{50,1}$	404.6	0.82	6.56	0.49	50.2	7.08
$\Gamma_{100,1}$	404.6	0.75	6.55	0.48	100.1	9.97
100	404.6	0.74	6.55	0.48	—	—
$\Gamma_{200,1}$	404.6	0.71	6.54	0.47	200.1	14.2
$\Gamma_{500,1}$	404.6	0.68	6.54	0.48	499.8	22.3
Parametric Gaussian	404.6	0.65	6.53	0.47	—	—

between θ and elements of this G^* vector ranged from -0.33 for G^*_1 to 0 for G^*_{100} to $+0.34$ for G^*_{200}, and correlations between σ and elements of the G^* vector ranged from $+0.67$ for G^*_1 to -0.06 for G^*_{100} to $+0.65$ for G^*_{200}.

The upshot of this inquiry is that if You know little, external to Your present dataset D, about the population empirical CDF G that gave rise to Your data (in a one-sample problem like that posed by the *NB10* dataset), and You express this uncertainty—in the Pólya-tree version of BNP modelling—with a prior on α that concentrates on small values (thereby ensuring that most of the information about G in the posterior comes from the empirical CDF based on Your sample), the resulting inferential answers for the parameters in Your model may not be well-calibrated, even if the centring distribution G_0 in Your Pólya-tree prior closely matches the actual data-generating mechanism. (Note that the conservatism in Table 20.3 is not present in the results summarized in Figure 20.5. I conjecture that this is because (a) θ was identified, in the modelling leading to Figure 20.5, as the median of G, whereas it was identified as the mean of G in the modelling that produced Table 20.3, and (b) the correlations noted in the previous paragraph are substantially smaller in the median modelling; this is a subject of continuing investigation.)

20.4 Bayesian nonparametric prior-distribution specification

Changing the focus now to specification of the prior distribution, it's common in Bayesian work to solve this specification problem with one member or another of a standard parametric family, sometimes chosen (e.g., for reasons of computational convenience) to be conjugate to Your sampling distribution. But this almost always goes beyond the optimal model-specification goal identified in Section 20.1; typically the sorts of propositions (relevant to Your prior distribution) that are rendered true by the context of the problem are (a) qualitative shape criteria such as monotonicity, convexity, or unimodality, and possibly also (b) one or more quantitative bounds on prior moments or percentiles. In such situations it would seem more satisfying to work with an infinite-dimensional non-parametric class $\mathcal{C}$ of prior densities satisfying the qualitative and quantitative criteria, for instance either (i) by sampling random members of this class and averaging over the implied uncertainty or (ii) by calculating upper and lower bounds over $\mathcal{C}$ for the posterior summaries of greatest interest (this is a form of sensitivity analysis).

Here's a case study in which to explore this idea. Suppose You're observing an IID Bernoulli(θ) process that has so far yielded n consecutive zeros, and the goal is to use the data to discriminate between two competing explanations for this outcome: $\theta = 0$ or $\theta > 0$. (In the real-world application on which this model is based, I once had occasion to buy n cups of tea over a several-week period from a machine that featured on its front a stick-on label announcing to customers that they might be lucky and get a free beverage, implying the existence of a device inside the machine that dispensed free drinks at random. After $n = 78$ consecutive fee-paying cups of tea, it was natural to speculate whether the makers of the machine had found it cheaper to attach the stick-on label, with no intent to offer free drinks at all, than to supply the machine with a randomization mechanism. Other applications of this problem arise, e.g., in medicine, when the first n patients screened in a particular sub-population all fail to have a disease that's rare in the overall population, and in process control, when the first n items manufactured have all been free of defects.)

Although most inferential settings involving observation of a Bernoulli process are more satisfyingly approached through interval estimation based on a model that treats $0 \leq \theta \leq 1$ continuously, with no individual value singled out for special treatment, this situation is a genuine sharp-null hypothesis-testing problem, and may be approached from the Bayesian point of view through the model

$$\left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & p(\theta|\mathcal{B}) \\ (y_i|\theta\ \mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \tag{20.11}$$

$(i = 1, \ldots, n)$, with a prior of the form

$$p(\theta|\mathcal{B}) = \left\{ \begin{array}{cc} 0 & \text{with probability } \lambda \\ \pi(\theta|\mathcal{B}) & (1-\lambda) \end{array} \right\} \tag{20.12}$$

for some $0 \leq \lambda \leq 1$. In the initial choice of $\pi(\theta|\mathcal{B})$ in the tea-machine case study, it seemed natural to quantify the following set of prior information about θ, conditional on θ being positive: (a) smaller values are more likely than bigger values, and (b) on substantive (economic) grounds, prior uncertainty about θ should be centred between two values (α_1, β_1), for instance $\left(\frac{1}{75}, \frac{1}{25}\right)$. (The upper bound in (b) arises because the makers of the tea machine would not wish to give away more

free drinks than necessary, and the lower bound corresponds to the view that, if θ were too small, customers would not perceive a large enough reward from the possibility of a free cup of tea for the randomization strategy to be worthwhile. Unimodal priors with a positive mode are also worth considering in this problem; this possibility will be examined elsewhere.)

20.4.1 A conjugate parametric solution

The off-the-shelf choice for $\pi(\theta|\mathcal{B})$ is of course a member of the Beta (η_1, η_0) family chosen, in view of (a) and (b), to be monotonically decreasing (and possibly also convex) and to have a mean between α_1 and β_1. Examination of the qualitative behaviour of the Beta family reveals that the desired monotonicity and convexity correspond to the region within which $0 < \eta_1 \leq 1$ and $\eta_0 \geq 2$. The mean constraint (b) in the Beta family,

$$\alpha_1 \leq \frac{\eta_1}{\eta_1 + \eta_0} \leq \beta_1, \tag{20.13}$$

further restricts the appropriate subclass of parametric priors to those with

$$\eta_1 \left(\frac{1}{\beta_1} - 1 \right) \leq \eta_0 \leq \eta_1 \left(\frac{1}{\alpha_1} - 1 \right), \tag{20.14}$$

giving rise with $\alpha_1 = \frac{1}{75}$ and $\beta_1 = \frac{1}{25}$ to the roughly triangular admissible region in Figure 20.6 (the contours in this plot will be explained below).

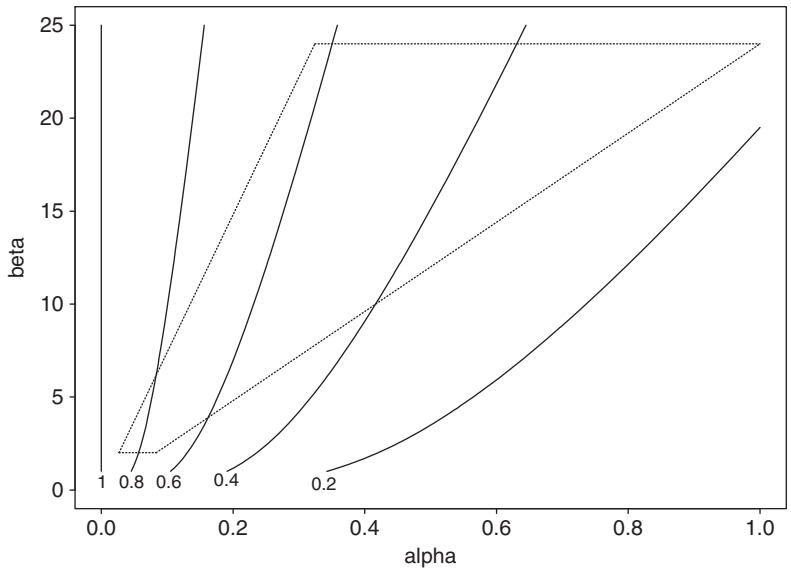

Figure 20.6 Admissible parametric priors given the substantive constraints (inside the dotted region), together with contours of $p(y|\theta > 0, \mathcal{B})$ with $n = 78$; see equation (20.16).

With data $y = (y_1, \ldots, y_n) = (0, \ldots, 0)$,

$$\begin{aligned}
\begin{bmatrix} \text{posterior} \\ \text{odds} \end{bmatrix} &= \begin{bmatrix} \text{prior} \\ \text{odds} \end{bmatrix} \cdot \begin{bmatrix} \text{Bayes} \\ \text{factor} \end{bmatrix} \\
\left[\frac{p(\theta = 0|y\,\mathcal{B})}{p(\theta > 0|y\,\mathcal{B})}\right] &= \left[\frac{p(\theta = 0|\mathcal{B})}{p(\theta > 0|\mathcal{B})}\right] \cdot \left[\frac{p(y|\theta = 0, \mathcal{B})}{p(y|\theta > 0, \mathcal{B})}\right] \\
&= \left(\frac{\lambda}{1-\lambda}\right) \cdot \left[\frac{1}{p(y|\theta > 0, \mathcal{B})}\right]
\end{aligned} \tag{20.15}$$

where

$$\begin{aligned}
p(y|\theta > 0, \mathcal{B}) &= \int_0^1 p(y|\theta, \theta > 0, \mathcal{B})\, p(\theta|\theta > 0, \mathcal{B})\, d\theta \\
&= \int_0^1 (1-\theta)^n\, \pi(\theta|\mathcal{B})\, d\theta \equiv B^{-1}.
\end{aligned} \tag{20.16}$$

With the parametric choice $\pi(\theta|\mathcal{B}) = \text{Beta}(\eta_1, \eta_0)$, the Bayes factor in favour of $\theta = 0$ takes the form

$$B(\eta_1, \eta_0) = \frac{\Gamma(\eta_0)\,\Gamma(\eta_1 + \eta_0 + n)}{\Gamma(\eta_1 + \eta_0)\,\Gamma(\eta_0 + n)}. \tag{20.17}$$

$0 < B^{-1}(\eta_1, \eta_0) < 1$ is a probability and is easier to contour-plot than the Bayes factor; the contours in Figure 20.6 are values of $B^{-1}(\eta_1, \eta_0)$ with $n = 78$. From this it may be seen that in the admissible region B^{-1} takes its minimum value 0.235 at $(\eta_1, \eta_0) = (1, 24)$ and its maximum value of 0.899 at $(\eta_1, \eta_0) = (0.027, 2)$. Thus in the parametric $\text{Beta}(\eta_1, \eta_0)$ class with the given prior specifications of monotonicity, convexity and bounds on the mean,

$$\frac{1}{0.899} = 1.11 \leq \begin{pmatrix} \text{Bayes factor} \\ \text{in favor of} \\ \theta = 0 \end{pmatrix} \leq 4.25 = \frac{1}{0.235}, \tag{20.18}$$

i.e., even with 78 consecutive zeros the strength of data evidence that $\theta = 0$ is surprisingly small. Using the informal guidelines of Jeffreys [15], as modified by Kass and Raftery [16], further calculation reveals that one would need more than 450 consecutive zeros for the evidence that $\theta = 0$ (as summarized by the upper bound on the Bayes factor) to pass from 'positive' to 'strong' with the prior specification examined here.

However, this conclusion is conditional on the Beta form of $\pi(\theta|\mathcal{B})$, which is not specified by the scientific context; how much bigger are the bounds when the calculation is made more appropriately over the nonparametric class $\mathcal{C}$ mentioned earlier? Answering this question involves finding the extreme values (here I mean supremum/infimum, which need not be attained) of the integral

$$I = I(\pi) = \int_0^1 (1-\theta)^n\, \pi(\theta|\mathcal{B})\, d\theta \tag{20.19}$$

when π ranges over $\mathcal{C}^*$, the set of functions $\pi(\theta|\mathcal{B})$: $[0,1] \to \Re$ in the constraint set

$$\left\{ \begin{array}{c} \pi(\theta|\mathcal{B}) \geq 0, \quad \int_0^1 \pi(\theta|\mathcal{B})\, d\theta = 1, \\ (*)\ \pi \text{ is monotone nonincreasing} \\ 0 < \alpha_1 \leq \int_0^1 \theta\, \pi(\theta|\mathcal{B})\, d\theta \leq \beta_1 \leq \frac{1}{2} \end{array} \right\}, \tag{20.20}$$

or the set $\mathcal{C}^{**}$ of $\pi(\theta|\mathcal{B})$ in the same constraint set but with $(*)$ replaced by

$$(**)\ \pi \text{ is monotone nonincreasing and convex.} \tag{20.21}$$

20.4.2 A nonparametric solution

Draper and Toland [6] give solutions to the nonparametric specification problems detailed in the previous paragraph, using a method based on functional analysis that appears to be new to the literature; space constraints here permit only a sketch of these results, itemized as follows.

- Let $\mathcal{C}^*$ and $\mathcal{C}^{**}$ be as in (20.20) and (20.21) for $n > 1$. Implementation of the method detailed in [6] leads to the conclusion that

$$\sup_{\pi \in \mathcal{C}^*} \int_0^1 (1-\theta)^n\, \pi(\theta|\mathcal{B})\, d\theta = 1 - \frac{2\,\alpha_1\, n}{n+1}. \tag{20.22}$$

- It turns out that this supremum is not attained by any $\pi \in \mathcal{C}^*$, but instead occurs at the generalized function

$$\pi^*_{\sup}(\theta|\mathcal{B}) = (1 - 2\,\alpha_1)\, \delta_0 + 2\,\alpha_1, \tag{20.23}$$

  where δ_0 is the Dirac delta measure at 0, i.e., the maximizing distribution has a point mass at 0 of size $(1 - 2\,\alpha_1)$ and is otherwise constant at height $2\,\alpha_1$ on $[0,1]$.
- One of the main ideas in [6] is to (a) identify a relaxed version of the optimization problem and then (b) relate the solutions of the relaxed problem to those of the original problem. To this end, Toland rewrites the primary problem (20.19) and (20.20) as follows:

$$\sup \,/\, \inf \left\{ 1 + \int_0^1 \left[(1-\theta)^n - 1 \right] \pi(\theta|\mathcal{B})\, d\theta \right\} \tag{20.24}$$

  over all functions $\pi : [0,1] \to \Re$ satisfying (20.20). This ensures that hypothesis (H_1) in Section 2.1 of [6] holds, and the relaxed problem is then (20.24) over all functions $\pi : [0,1] \to \Re$ in the relaxed constraint set

$$\left\{ \begin{array}{c} \pi(\theta|\mathcal{B}) \geq 0, \quad \int_0^1 \pi(\theta|\mathcal{B})\, d\theta \leq 1, \\ \pi \text{ is monotone nonincreasing} \\ 0 < \alpha_1 \leq \int_0^1 \theta\, \pi(\theta|\mathcal{B})\, d\theta \leq \beta_1 \leq \frac{1}{2} \end{array} \right\}. \tag{20.25}$$

Table 20.4 Bayes factor bounds as a function of how the prior is specified, with $n = 78, \alpha_1 = \frac{1}{75}$, and $\beta_1 = \frac{1}{25}$.

	Bayes factor	
Specification	**Low**	**High**
Parametric	1.11	4.25
Nonparametric $\mathcal{C}^*$	1.03	6.33
Nonparametric $\mathcal{C}^{**}$	1.03	5.29

The main point of the discussion in this case study is the observation that the supremum and infimum of the relaxed problem *are attained* at the constant function $\pi \equiv 2\alpha_1$, and coincide with the supremum and infimum of the primary problem (20.19, 20.20) (even though the latter supremum is *not attained*).

- The infimum of I over $\pi \in \mathcal{C}^*$ turns out to be

$$\inf_{\pi \in \mathcal{C}^*} \int_0^1 (1-\theta)^n \, \pi(\theta|\mathcal{B}) \, d\theta = \frac{1-(1-2\beta_1)^{n+1}}{2\beta_1 (n+1)}; \tag{20.26}$$

the infimum is attained in $\mathcal{C}^*$ by

$$\pi^*_{\text{inf}}(\theta|\mathcal{B}) = \left\{ \begin{array}{cc} \frac{1}{2\beta_1} & \text{for } 0 \le \theta \le 2\beta_1 \\ 0 & 2\beta_1 < \theta \le 1 \end{array} \right\}, \tag{20.27}$$

i.e., a piecewise constant density (histogram).

- When convexity is added the supremum is unchanged, but the infimum of I over $\mathcal{C}^{**}$ is

$$2\left[\frac{1}{3\beta_1 (n+1)} - \frac{1-(1-3\beta_1)^{n+2}}{(3\beta_1)^2 (n+1)(n+2)}\right] \tag{20.28}$$

and occurs at the function

$$\pi^{**}_{\text{inf}}(\theta|\mathcal{B}) = \left\{ \begin{array}{cc} \frac{2}{(3\beta_1)^2}(3\beta_1 - \theta) & \text{for } 0 \le \theta \le 3\beta_1 \\ 0 & 3\beta_1 < \theta \le 1 \end{array} \right\}, \tag{20.29}$$

i.e., a piecewise linear density (a frequency polygon).

With $n = 78, \alpha_1 = \frac{1}{75}$, and $\beta_1 = \frac{1}{25}$, the minimum and maximum values of I over $\mathcal{C}^*$ are 0.158 and 0.974, respectively, and over $\mathcal{C}^{**}$ the minimum rises to 0.189. Table 20.4 summarizes the numerical findings, and Figure 20.7 plots the optimizing densities. Without convexity the nonparametric limits are 8% lower and 49% higher than the parametric values, and even with convexity the corresponding figures are 8% and 24%; the casual adoption of a convenient parametric family satisfying the scientifically motivated monotonicity and convexity constraints has led to noticeably narrower sensitivity bounds than those that more appropriately arise from assuming only monotonicity and convexity.

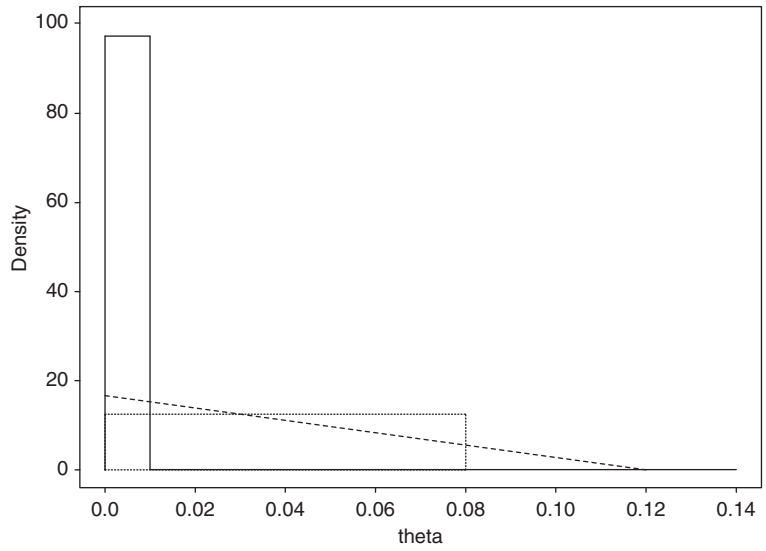

Figure 20.7 Optimal densities π^*_{inf} (long dotted line), π^{**}_{inf} (short dotted line), and approximation to π^*_{sup} in which the point mass at 0 is replaced by a histogram bar of width $\epsilon = 0.01$ (solid line).

20.4.3 A heuristic sketch of the method of proof

The goal is to optimize the integral $I = I(\pi)$ in (20.19) over functions π subject to monotonicity and/or convexity constraints, (20.20) and (20.21), respectively. The basic idea is to think of π as though it were a point in a convex subset of $\Re^k$ and to use intuitions from the geometry of such sets in $\Re^k$ to inform the proof.

(1) The set $\mathcal{C}^*$ of admissible functions π may be thought of as like a convex polygon in $\Re^k$, but with infinitely many vertices and edges. In this heuristic proof sketch, consider the situation with $k = 2$ and visualize a polygon with a finite number of vertices and edges. The boundary points are of two kinds: extreme points and all the other boundary points. An extreme point is not an interior point of any line segment in the set. (For example, the corners of a square are its only extreme points.)

(2) The function I in (20.19) is *linear* on π, analogous to a linear function in the plane. But it's important to realize that, because the set of functions satisfying (20.19) and (20.20) is infinite-dimensional, the extreme values may not be attained at points of the constraint set; and in fact the supremum is not attained but the infimum is. Because of this difficulty, Toland replaced the primary problem with a relaxed problem over a larger convex set and showed that in the relaxed problem the extreme values are attained and that they coincide with the extreme values of the original problem. Moreover he showed that the solution of the relaxed problem is finite-dimensional, although the original problem was not.

(3) The level sets of a linear function on $\Re^2$ are parallel straight lines. When the relaxed polygon lies on one side of a level line and touches it, the polygon and the line intersect in one of two ways: at single vertices (one for each of the minimizer and maximizer), or along an edge. In the latter case, even though the intersection set includes points that are not vertices, the set always includes vertices. Thus it suffices in optimizing a linear function to evaluate it at

the vertices of the polygon (this is the basic intuition behind linear programming). This is where the *Krein–Milman theorem* (e.g., [21]) comes into the infinite-dimensional argument (see Theorem 2 in Section 2 of [6]).

(4) Because of geometric constraints of monotonicity on π in the relaxed problem, the analogue of vertices in $\mathcal{C}^*$ turns out to be the class of *step functions* with at most three distinct values; thus the infinite set of extrema can be indexed with at most five parameters.

(5) A vertex π of the relaxed polygon has the property that no linear perturbation ($\pi \pm \phi\,\pi$) away from it for small non-zero ϕ lies completely inside the admissible set. Toland was able to conjecture a particular form of ϕ, namely

$$\phi(\theta) = \int_0^\theta h(t)\,\pi'(t)\,dt \tag{20.30}$$

and show that if π is more complicated than a step function with two distinct values, a non-trivial ($\phi\,\pi \neq 0$) h can always be found such that all of the constraints in $\mathcal{C}^*$ are satisfied. Therefore the vertices are histograms with two or three bars.

(6) Thinking of a two-bar histogram, in the limit as the left-hand bar becomes infinitely tall the maximizer π^*_{sup} over $\mathcal{C}^*$ results: a point mass at 0 plus a constant over the rest of $[0, 1]$. In the limit as the right-hand bar goes to 0 the minimizer over $\mathcal{C}^*$ is obtained.

(7) When convexity is added, the relaxed polygon vertices become the class of *convex piecewise linear functions* with exactly three distinct segments; here only six parameters are needed. The maximizer over $\mathcal{C}^{**}$ remains the same as in $\mathcal{C}^*$ because π^*_{sup} is already convex. The minimizer of the relaxed problem turns out to be a two–part piecewise linear function (frequency polygon) with the second segment 0.

(8) When unimodality is assumed instead of the other qualitative constraints examined, a modification of the method Toland employs for dealing with monotonicity is available, because unimodal densities on $[0, 1]$ are non-decreasing on $[0, d)$ and non-increasing on $(d, 1]$ for some $d \in [0, 1]$.

Results similar to the findings here under monotonicity have been obtained elsewhere in the Bayesian robustness literature by quite different means, through the use of Khintchine's Theorem (e.g., [8]) on generating unimodal densities as mixtures of uniform distributions. The approach sketched here, via functional analysis, both subsumes the condition of unimodality and yields new results under the alternative qualitative specifications of monotonicity and convexity.

20.4.4 Conclusions

As more moment constraints are added to the quantitative mean constraint examined here to increase realism, e.g.,

$$\sigma^2_{low} \leq \int_0^1 [\theta - E(\theta|\mathcal{B})]^2\ \pi(\theta|\mathcal{B})\,d\theta \leq \sigma^2_{high}, \tag{20.31}$$

the optimal nonparametric solutions become k–part piecewise linear functions (frequency polygons) with increasing k, approaching the smoothness built into parametric families like Beta (α, β). Thus continuous parametric assumptions are equivalent to infinite sets of moment constraints, i.e., when You choose continuous parametric priors You're probably assuming more than You think You are.

The set of practical problems in which

(a) the prior really matters and
(b) off-the-shelf parametric specifications are often used instead of qualitative descriptions involving shape (e.g., number of modes, monotonicity, convexity, smoothness) and substantive bounds on quantitative descriptions (e.g., moments or quantiles)

is larger than is generally acknowledged. The method of proof offered here shows promise to inform Bayesian sensitivity analysis in a wide variety of such problems, because all of the above characteristics may be enforced with linear constraints through derivatives and integrals of π.

Postscript On the 79th visit to the machine it yielded a free cup of tea.

Acknowledgements

I'm grateful to Tim Hanson and Alejandro Jara (a) for help with the programming that led to the results presented in Section 20.3 and (b) for comments that aided interpretation of the Bayesian nonparametric findings; to Milovan Krnjajić for helpful discussions about calibration cross-validation; to John Toland for doing all of the heavy lifting in the proofs underlying Section 20.4; and to Jim Berger, Brad Efron, Richard Olshen, Luc Tartar, and Stephen Walker for helpful comments and references. Membership on this list does not imply agreement with the conclusions drawn here, nor are any of these people responsible for any errors that may be present.

References

[1] Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, **14**, 1–13.
[2] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1–68.
[3] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, **57**, 45–97.
[4] Draper, D. (2012). Bayesian model specification: towards a theory of applied statistics. Submitted.
[5] Draper, D. and Krnjajić, M. (2012). Calibration results for Bayesian model specification. Submitted.
[6] Draper, D. and Toland, J. (2012). Bayesian non-parametric prior specification. Technical Report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.
[7] Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.
[8] Feller, W. (1971). *An Introduction to Probability Theory and its Applications* (2nd edn), Volume 2. Wiley, New York.
[9] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 209–230.
[10] Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics*, **34**, 1194–1216.
[11] Freedman, D., Pisani, R. and Purves, R. (2007). *Statistics* (4th edn). Norton, New York.

[12] Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.
[13] Hanson, T. (2006). Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, **101**, 1548–1565.
[14] Jara, A., Hanson, T., Quintana, F., Müller, P. and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, **40**, 1–30.
[15] Jeffreys, H. (1967). *Theory of Probability* (3rd edn). Oxford University Press, Oxford.
[16] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
[17] Krnjajić, M., Kottas, A. and Draper, D. (2008). Parametric and non-parametric Bayesian model specification: a case study involving models for count data. *Computational Statistics and Data Analysis*, **52**, 2110–2128.
[18] Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1222–1235.
[19] Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.
[20] Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer-Verlag, New York.
[21] Rudin, W. (1991). *Functional Analysis* (2nd edn). McGraw-Hill, New York.