Chapter 1

Formal Bayes Methods for Model Calibration with Uncertainty*

This box describes the Bayesian approach to assessing uncertainty, and how it can be implemented to calibrate model parameters using observations, taking account of the imperfections of the model, and measurement errors. Section 1.1 outlines the justification for the Bayesian approach, Sec. 1.2 outlines the Bayesian approach to model calibration, and sections 1.3 and 1.4 discuss simple and more advanced strategies for performing the inferential calculations. There is a brief summary in Sec. 1.5.

1.1. Bayesian methods

Bayesian methods provide a formal way of accounting for uncertainty, through the use of probability, and the probability calculus. Uncertainty, treated generally, is a property of the mind; it pertains to an individual, and to the knowledge that individual possesses. Many people baulk at the uncompromisingly subjective or 'personalistic' nature of uncertainty. A superficial understanding of science would suggest that this subjectivity is out of place, but in fact it lies at the very heart of what makes a scientist an expert in his or her field: the capacity to make informed judgements in the presence of uncertainty (Ziman, 2000, provides a naturalistic assessment of 'real' science). Different Hydrologists will produce different models of the same catchment, which is to say that the process of designing and constructing a model is subjective. The Bayesian approach extends this subjectivity to descriptions of uncertainty, e.g. uncertainty about the relationship between the model output and the behaviour of the actual catchment. But while model-building is something Hydrologists do a lot of, thinking about uncertainty is less familiar, and seems less structured.

*Author: Jonathan Rougier, Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW. Email j.c.rougier@bristol.ac.uk. A contribution to K. Beven and J. Hall (eds), *Applied Uncertainty Analysis for Flood Risk Management*, Imperial College Press / World Scientific, due out 2008.

And yet it is a vital part of any model-based analysis—we cannot make inferences about a catchment without accounting for the limitations of the model. The probabilistic approach is therefore a way of making explicit what must be happening implicitly. In requiring us to quantify our uncertainties as probability distributions, it puts these judgements into a form where they may be debated, and amended (Goldstein, 2006).

The fact that these judgements are subjective, and the case for making them transparent in scientific inference, are unassailable. What we have yet to establish here is why we should do this within a probabilistic framework. The pragmatic answer is that the probabilistic approach has proved to be extremely powerful and, in conjunction with modern computational methods (particularly Monte Carlo methods), is unsurpassed in complex inferences such as data assimilation, spatial-temporal modelling, and scientific model calibration and model-based prediction (see also the many scientific applications in Liu, 2001). As these fields have developed, a consensus has emerged, and the result is that the overt subjectivity has been somewhat reduced, in the same way that a consensus on how to treat a certain aspect of a hydrologic model reduces the differences across models.

The pragmatic answer focuses on the efficacy of the probability calculus. Perhaps that is the only justification that is required. Before the advent of modern computational methods, though, the first answer would have been that there is foundational support for the probability calculus as a model for the way we reason. The probability calculus is based on three simple axioms. We suppose the existence of a set Ω, and a measure $\Pr(\cdot)$ defined on subsets of Ω. The axioms assert that $\Pr(\cdot)$ satisfies the following properties:

(1) $\Pr(A) \geq 0$;
(2) $\Pr(\Omega) = 1$;
(3) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \emptyset$;

where $A, B \subseteq \Omega$ (see, e.g., Dawid, 1994). Why these axioms and not others? There are a number of interpretations, i.e., suggested relations between these axioms and the real world (see, e.g., Gillies, 1994). In the Bayesian interpretation, $\Pr(A)$ is an operationally-defined assessment of an individual's uncertainty about A, a view that was formalised by Bruno de Finetti in the 1930s (de Finetti, 1937, 1964, 1972). Book-length treatments of this approach can be found in Savage (1972) and Lad (1996); Jeffrey (2004) provides a concise introduction.

Imagine that we are interested in X, the amount of rain in mm

on the Met Office roof on Christmas Day 1966. We might set $\Omega = \{X = 0, X = 1, \ldots, X = 100\}$. Any subset of Ω is termed a proposition, and interpreted as the union of its components. Thus if $A = \{X = 0, X = 1, X = 2\}$ then $\Pr(A) = \Pr(X = 0 \text{ or } X = 1 \text{ or } X = 2)$. But how does one assess $\Pr(A)$? One operationalisation of the Bayesian approach is to think of $\Pr(A)$ as the value of v that minimises, for me (or whoever's probability is being assessed), the loss function $(v - I_A)^2$, where I_A is the indicator function of the proposition A, i.e. $I_A = 1$ if A is true, and 0 otherwise. If I was sitting at my desk with the meteorological records for the Met Office roof in front of me, I would know whether or not A was true. In this case $v = 1$ would minimise my loss if it was, and $v = 0$ if it was not. In general, however, I would likely settle on a value for v somewhere between 0 and 1: where exactly would be a quantification of how probable I thought that A was.

The operational definition of probability is combined with a simple rationality principle: I would never choose a probability (or, more generally, collection of probabilities) which resulted in a loss that could be unambiguously reduced no matter what the outcome. Probabilities obeying this principle are termed *coherent*. It is easy to show that coherence implies the three axioms given above. For example, if I chose a value for $\Pr(A)$ that was strictly less than zero, then a value $\Pr(A) = 0$ would result in a loss that was smaller, no matter whether A turned out to be true or false; hence $\Pr(A) \geq 0$ is implied by coherence.

In order for these axioms to lead to a useful calculus, we need rule for describing how knowing the truth of one proposition would change our probabilities for others. In other interpretations of the probability axioms this is defined to be the *conditional* probability $\Pr(A \mid B) = \Pr(A \cup B) / \Pr(B)$, provided that $\Pr(B) > 0$. In the Bayesian approach, however, the conditional probability $\Pr(A \mid B)$ is operationally defined as the value of v which minimises, for me, the loss function

$$I_B(v - I_A)^2 \tag{1.1}$$

a definition which subsumes $\Pr(A)$, which is equal to $\Pr(A \mid \Omega)$ since $I_\Omega = 1$ with certainty. It can then be proved that the relation

$$\Pr(A \cup B) = \Pr(A \mid B) \Pr(B) \tag{1.2}$$

follows as a consequence of the coherence of the collection of probabilities

for $A \cup B$, $A \mid B$, and B (see, e.g., de Finetti, 1972, ch. 2). The result

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \qquad \text{providing that } \Pr(B) > 0 \tag{1.3}$$

which is an immediate consequence of (1.2), is referred to as Bayes's Theorem precisely because it is a theorem: it is a consequence of the operational definition of $\Pr(A \mid B)$ and the principle of coherence.

The Bayesian approach does not assert that this is how people actually assess probabilities: it is a model for reasoning, and has the same advantages and disadvantages as models used elsewhere in science. For simple propositions we can usually assess $\Pr(A)$ directly, without recourse to thinking about loss functions: most people seem to understand probability without having to operationalise it. For more complicated propositions, however, the probability calculus helps us to break probability assessments down into more manageable parts.

Suppose, for example, that X denotes uncertain parameters in our hydrologic model, and z^{obs} a sequence of flow measurements on a river downstream of the catchment (conventionally, capitals are used for uncertain quantities, and small letters for possible values). We *might* attempt to describe $\Pr(X = x)$ for all x directly, taking into account only *implicitly* that $Z = z^{\text{obs}}$. It is important to stress this point: $Z = z^{\text{obs}}$ *is* known at the point where $\Pr(X = x)$ is specified, and so if this information is relevant it ought to impact on $\Pr(X = x)$; but the impact is implicit, in that we cannot trace, formally, the effect that Z equals z^{obs} rather than some other value.

Bayes's Theorem gives us an alternative way of assessing X: we can compute the *posterior* distribution

$$\Pr\big(X = x \mid Z = z^{\text{obs}}\big) = \frac{\Pr\big(Z = z^{\text{obs}} \mid X = x\big)\Pr(X = x)}{\Pr(Z = z^{\text{obs}})}. \tag{1.4}$$

In this case, the link between $Z = z^{\text{obs}}$ and $X = x$ is made explicit. The cost, though, is that now we have to specify two distributions instead of one: we need to specify the conditional distribution $\Pr(Z = z \mid X = x)$ for all values z and x, and the *prior* distribution $\Pr(X = x)$ for all x, then we can compute the required conditional probability (the denominator is simply a normalising constant). The first function in the numerator of Eq. (1.4),

$$L(x) \triangleq \Pr\big(Z = z^{\text{obs}} \mid X = x\big)\,, \tag{1.5}$$

is known as the *likelihood function*, where '$\triangleq$' denotes 'defined as'. The likelihood function is a function of just x, and it appears as though $\Pr(Z = z \mid X = x)$ is not required, except where $z = z^{\text{obs}}$. But the validity of Bayes's Theorem depends on the likelihood function being one particular value from a well-defined conditional distribution, and so we have to specify the whole distribution $\Pr(Z = z \mid X = x)$, and then plug-in z^{obs}.

Other methods for uncertainty assessment, such as the GLUE approach (see Box ??? in this volume), specify a 'likelihood-like' function: a function of x for which smaller values indicate a poorer fit to the data. Inferences based on these 'likelihood-like' functions cannot be formally interpreted as probabilities. They might, however, be informally interpreted as probabilities: within the subjective framework there is nothing to stop an individual from adopting whatever method he or she sees fit to assess his or her probabilities. The issue, though, is whether the resulting assessments are authoritative. There is, perhaps, a lack of authority in a probabilistic assessment that cannot be demonstrated to be consistent with the probability calculus.

Most of us make our everyday probabilistic assessments directly. For example, when we assess Pr(rain today) we take account, informally, of the event 'rain yesterday': we do not do the conditional probability calculation. In scientific applications, though, the conditional calculation has a large advantage: the conditional distribution $\Pr\big(Z = z^{\text{obs}} \mid X = x\big)$ is often 'nearly' available, in the form of a physical model. Denoting Y as the actual behaviour of the river, the model represents a function mapping candidate values for X into candidate values for Y: the 'forwards' direction of the model corresponds to the direction of the conditional probability $\Pr(Z = z \mid X = x)$. Of course the model does not get us all the way from X to Z, because we still have to account for the effect of its inaccuracies, and of measurement errors. But it is reasonable to expect the model, if it is carefully constructed, to get us most of the way there. Therefore while there is nothing to stop us trying to assess $\Pr(X = x)$ directly, taking account informally that $Z = z^{\text{obs}}$, in scientific applications there is often a strong case for using the conditional probability $\Pr\big(X = x \mid Z = z^{\text{obs}}\big)$ instead, in conjunction with a physical model.

The next section describes a statistical framework for linking together the model parameters, model evaluations, the actual system values, and the observations.

1.2. Model calibration in a simple statistical framework

As above, let X denote the (unknown) 'best' values for the model parameters, let Y denote the system values, e.g. river height at various locations and various times, let Z denote the measurements, and z^{obs} the actual measured values. *Calibration* is learning about X using $Z = z^{\text{obs}}$; *calibrated prediction* is learning about X and Y using $Z = z^{\text{obs}}$. This Box focuses on calibration, but the extension to calibrated prediction is straightforward. The main purpose of calibration is to assess a point value and a measure of uncertainty for the 'best' values for the model parameters. Craig et al. (1997, 2001) and Goldstein and Rougier (2006) discuss the statistical approach to calibration and calibrated prediction, particularly for large problems; Kennedy and O'Hagan (2001) provide a more conventional but less scalable approach. Note that the assertion that there exists a 'best' value for the model parameters is not clear-cut; this is discussed within the context of a more general statistical framework in Goldstein and Rougier (2004, 2007).

Referring back to Eq. (1.4), we need to specify the prior distribution $\Pr(X = x)$ for each x, and the statistical model $\Pr(Z = z \mid X = x)$, for each combination of x and z. The prior distribution quantifies our judgements about the model parameters before observing Z (or, more realistically, neglecting the information that $Z = z^{\text{obs}}$). It must respect the physical limitations of the model parameters, but it should also reflect information collected from previous observations on the catchment, or similar catchments, where such information exists. Sometimes a fairly vague specification for $\Pr(X = x)$ will suffice, in situations where there is lots of information about X in Z. The 'classical' situation where this occurs is where the components of Z are independent conditional on X. However, this is emphatically not the case with physical models, for reasons to be explained below. Therefore the choice of $\Pr(X = x)$ is likely to have some impact on the posterior distribution, and it is worth investing some effort in this choice or, if that is not possible or if it proves too hard, performing a sensitivity analysis by re-doing the inference for a range of choices.

We also have to specify the statistical model $\Pr(Z = z \mid X = x)$. We could specify this directly, but in practice it is easier to induce this distribution by specifying two other quantities. Denote the model output at model parameters x as $f(x)$. We will assume, for simplicity, that the components of $f(x)$, Y, and Z correspond one-to-one, for every x. The difference $Y - f(x)$ denotes the difference between the system values and the model

output when evaluated at x. This is uncertain because Y is uncertain. The difference

$$\epsilon \triangleq Y - f(X) \tag{1.6}$$

denotes the *model discrepancy*. This is the difference between the system value and the model output when evaluated at the best choice of parameter values, X. This is uncertain because both X and Y are uncertain. Next, we need a statistical model for the *measurement error*, to take us from Y to Z,

$$e \triangleq Z - Y. \tag{1.7}$$

Putting these together, we have a statistical model for the distribution of Z conditional on X, since

$$Z \equiv Y + e \equiv f(X) + \epsilon + e, \tag{1.8}$$

where '$\equiv$' denotes 'equivalent by definition'.

In the simplest case where X, ϵ and e are treated as mutually independent, a treatment that is almost always used in practice, our choices for the marginal distributions of ϵ and e induce the conditional distribution $\Pr(Z = z \mid X = x)$. For example, suppose that we decide that both ϵ and e are multivariate Gaussian (this might require a transformation of $f(x)$, Y, and Z), each with mean zero, and with variance matrices Σ^ϵ and Σ^e. Exploiting the fact that the sum of two independent Gaussian distributions is Gaussian, we find that

$$L(x) = \varphi\big(z^{\text{obs}}; f(x), \Sigma^\epsilon + \Sigma^e\big) \tag{1.9}$$

where $\varphi(\cdot)$ is the Gaussian Probability Density Function (PDF) with specified mean and variance. We will adopt these choices from now on, so that our task simplifies to (i) choosing a prior distribution for X and choosing the variance matrices Σ^ϵ, and Σ^e, and (ii) calculating $\Pr\big(X = x \mid Z = z^{\text{obs}}\big)$ on the basis of these choices. Strategies for doing the calculation are discussed in sections 1.3 and 1.4.

Now we can clarify why physical models do not give rise to observations that are conditionally independent given X. When physical models are inaccurate, their errors are almost always systematic across the output components. For example, if the model predicts a value that is too high at time t, then we would usually judge that this error will persist into time $t + 1$, if the unit of time is not too large. This persistence of errors is represented by dependence among the components of the discrepancy ϵ. This

dependence means that the observations are not conditionally independent given X. The only situation in which we can ignore the discrepancy (in the sense that it has little effect on the inference) is when it is dominated by the measurement error, so that $\Sigma^\epsilon + \Sigma^e \approx \Sigma^e$. It is fairly standard to treat the measurement errors as independent, and in this case the observations would be conditionally independent given X. But if the observation errors are large, then the data are not very informative about X, and so the choice of prior $\Pr(X = x)$ will be important.

Specifying the discrepancy variance Σ^ϵ is the hardest task in calibrating a model. Often it is ignored, i.e. implicitly set to zero. In this case each observation is treated as more informative than it actually is, and the result can be incompatible posterior distributions based on different subsets of the observations. Another example of a poor implicit choice is to minimise the sum of squared differences between z^{obs} and $f(x)$. This is equivalent to finding the mode of the posterior distribution in the special case where both Σ^ϵ and Σ^e are treated as proportional to the identity matrix. The problem with this choice is that it ignores the persistence of model errors, and so over-weights collections of observations that are close in space or time. A crude way around this is to thin the observations, arranging that they are sufficiently well-separated that the persistence is negligible. This is an effective strategy if the observations are plentiful, and it reduces the specification of Σ^ϵ to a diagonal matrix: perhaps even simply $\sigma_\epsilon^2 I$ for some scalar σ_ϵ and identity matrix I, if all the observations are the same type. In general, however, the diagonal components of Σ^ϵ will have to be set according to how good the model is judged to be, and the off-diagonal components according to how persistent the model-errors are judged to be.

Rougier (2007) discusses these issues in more detail, in the context of climate modelling.

1.3. Simple sampling strategies

The posterior distribution in Eq. (1.4) is very unlikely to have a closed-form solution (which would only happen if the model was linear and the prior $\Pr(X = x)$ was Gaussian). Therefore either we estimate the constant of integration, $\Pr\big(Z = z^{\text{obs}}\big)$, or we use a random sampling scheme that does not require this value to be computed explicitly. For simplicity, we will assume from now on that X is absolutely continuous with prior PDF

$\pi_X(x)$ and posterior PDF $\pi_{X|Z}(x)$, for which Bayes's Theorem states

$$\pi_{X|Z}(x) = c^{-1} L(x)\, \pi_X(x) \quad \text{where} \quad c \triangleq \int_{\mathcal{X}} L(x)\, \pi_X(x)\, \mathrm{d}x, \tag{1.10}$$

where $\mathcal{X} \subseteq \mathbb{R}^p$ is the parameter space, and c is the *normalising constant* (also known as the *marginal likelihood*), which we previously denoted as $\Pr\big(Z = z^{\mathrm{obs}}\big)$.

If p (the dimension of $\mathcal{X}$) is low, say less than five, the former approach may be the best option. In this case, a deterministic numerical scheme can be used to approximate c (see, e.g., Davis and Rabinowitz, 1984; Kythe and Schäferkotter, 2004). Once this value has been computed, the posterior distribution can be summarised in terms of means, standard deviations and correlations, using further integrations. If a single point-estimate of X is required, the posterior mean is usually a good choice.

If p is much larger than about five, though, this approach becomes unwieldy, because so many points are required in the integration grid. The alternative strategy is to randomly sample from the posterior directly, which can then be summarised in terms of the properties of the sample. These properties include quantiles, since the empirical distribution function of any parameter can be computed directly from the sample, and this can then be inverted. Sampling does not of itself fix the problem of a high-dimensional parameter space. In particular, n function evaluations in a random sampling scheme are likely to do a worse job than n points in an integration scheme, since in the latter these points will be chosen to span $\mathcal{X}$. But the overriding advantage of sampling is its flexibility: we can keep going until the summaries of the posterior are accurate enough for our purposes, and we can adapt our approach as we go along. Numerical integration requires us to operate on pre-specified grids, and if we find out that an n-point grid does not deliver the required accuracy, then it is hard to reuse these points in a more accurate calculation on a new, denser, grid (although Romberg integration is one possibility, see Kythe and Schäferkotter, 2004, sec. 2.7).

The subject of Monte Carlo sampling is huge and still developing (see, e.g., Ripley, 1987; Robert and Casella, 1999; Evans and Swartz, 2000; Liu, 2001). Here we outline one of the simplest approaches, *Importance Sampling*, since it is intuitive and corresponds quite closely to current practice (see the end of this section). The mantra for the most basic form of Importance Sampling is *sample from the prior, weight by the likelihood*. The following steps are repeated for $i = 1, \ldots, n$:

(1) Sample $x^{(i)}$ from $\pi_X(x)$;
(2) Evaluate the model to compute $f(x^{(i)})$;
(3) Now compute the weight $w_i \triangleq L(x^{(i)})$, e.g. using Eq. (1.9).

The weights describe the quality of the fit between each $f(x^{(i)})$ and z^{obs}, taking account both of the model discrepancy and the observation error. Upweighting candidates for X that give a good fit to the observations is very intuitive, but this general principle gives no guidance regarding the form of the weighting function. The Bayesian formalism indicates that in this simple approach (a more sophisticated approach is described in Sec. 1.4) the correct choice for the weighting function is the likelihood function.

After n samples, the mean of the weights is an estimate of c:

$$c \equiv E\big(L(X)\big) \approx n^{-1}\big(w_1 + \cdots + w_n\big), \tag{1.11}$$

where the expectation is with respect to $\pi_X(x)$. To estimate the posterior expectation of some specified function $h(X)$ we compute

$$\begin{aligned} E\big(h(X)\big) &= c^{-1} \int_X h(x)\, L(x)\, \pi_X(x)\, \mathrm{d}x \\ &= c^{-1} E\big(h(X)\, L(X)\big) \\ &\approx \frac{w_1\, h(x^{(1)}) + \cdots + w_n\, h(x^{(n)})}{w_1 + \cdots + w_n}, \end{aligned} \tag{1.12}$$

where n^{-1} cancels top and bottom. Thus to estimate the mean vector μ we choose $h(x) = x$, and to estimate the variance of X_i or the covariance between X_i and X_j we choose $h(x) = (x_i - \mu_i)(x_j - \mu_j)$, where for the variance $j = i$.

We can also estimate quantiles, by inverting the distribution function. The cumulative probability $\Pr\big(X \le x' \mid Z = z^{\text{obs}}\big)$ can be estimated for any x' by setting $h(x) = I_{x \le x'}$, remembering that $I_{x \le x'}$ is the indicator function. For simplicity, suppose that X is a scalar (i.e., $p = 1$). Then the estimated posterior distribution function of X has steps of

$$\frac{w_{(1)} + \cdots + w_{(i)}}{w_1 + \cdots + w_n} \tag{1.13}$$

at each $o_{(i)}$, where $o_{(1)}, \ldots, o_{(n)}$ are the ordered values of $x^{(1)}, \ldots, x^{(n)}$, and $w_{(1)}, \ldots, w_{(n)}$ are the correspondingly-ordered weights. This distribution function can be inverted to give marginal posterior quantiles; i.e. we identify that value $o^{(i)}$ for which $\Pr\big(X \le o^{(i)} \mid Z = z^{\text{obs}}\big)$ is approximately equal to our target probability. An intuitive measure of uncertainty about X is the 95% *symmetric credible interval*, which is defined by the 2.5th and 97.5th

percentiles. It is a mistake often made in practice, but this should *not* be referred to as at 95% *confidence interval*, which is quite a different thing (see, e.g., a standard statistics textbook such as DeGroot and Schervish, 2002, sec. 7.5).

Comparison with current practice. Current practice for model-calibration is diverse, but one strategy that is used frequently is to sample $x^{(i)}$ from the PDF $\pi_X(x)$ and then keep all those samples for which some distance measure on $z^{\text{obs}} - f(x^{(i)})$ is small, discarding the rest. This cannot be consistent with the Importance Sampling approach outlined in this section unless the likelihood function is zero for some set of parameter values, and constant on the complement of this set in $\mathcal{X}$. It is hard to imagine such a likelihood function emerging from any reasonable choice for the conditional probability of Z given X; hence, this approach cannot be said to give rise to a sample from the posterior PDF $\pi_{X|Z}(x)$.

This strategy of keeping only those samples that match the observations sufficiently well requires us to quantify what we mean by 'sufficiently well', and any number of different choices are possible, although the Nash-Sutcliffe measure seems to be the most popular in Hydrology. This issue is resolved in the Importance Sampling approach, which tells us exactly how the difference between z^{obs} and $f(x^{(i)})$ should be scored if we want to describe our uncertainty probabilistically. But the major benefit of Importance Sampling arises from its efficiency in more advanced implementations, as discussed in the next section.

1.4. More advanced strategies

Sampling from the prior and weighting by the likelihood is very intuitive. It works well in situations where the observational data are not highly informative, so that the posterior PDF $\pi_{X|Z}(x)$ is not that different from the prior, $\pi_X(x)$. This is because the sampled values $\left\{x^{(1)}, \ldots, x^{(n)}\right\}$ do a good job of spanning the x-values that predominate in the posterior. Typically the observational data will be not-highly-informative when the measurement errors are large, when the discrepancy is large (i.e. the model is judged to be poor), or when the model output is fairly constant in x.

What about the other situation, though, when the observations are expected to be highly informative. In this case, simple Importance Sampling 'wastes' model evaluations by putting many of the $x^{(i)}$ into regions of the parameter space that have near-zero likelihood, and thus near-zero poste-

rior probability. This will be a problem if the model itself is expensive to evaluate. In this case, it would be more efficient to find a way to sample the $x^{(i)}$ so that they were likely to occur in regions of high posterior probability. Importance Sampling allows us to do this, and to correct for the fact that we are not sampling from $\pi_X(x)$.

Suppose we think that some specified PDF $\pi'_X(x)$ is likely to be a better approximation to the posterior PDF than is $\pi_X(x)$, and that $\pi'_X(x)$ is easy to sample from and to compute; $\pi'_X(x)$ is known as the *proposal* distribution. The PDFs $\pi_X(x)$ and $\pi'_X(x)$ must satisfy certain technical conditions: $\pi'_X(x) > 0$ wherever $\pi_X(x) > 0$, and the ratio $L(x)\pi_X(x)/\pi'_X(x)$ must be strictly bounded above (these are discussed further below). Where these conditions hold, the sampling strategy is:

(1) Sample $x^{(i)}$ from $\pi'_X(x)$;
(2) Evaluate the model to compute $f(x^{(i)})$;
(3) Now compute $w_i \triangleq L(x^{(i)})\,\pi_X(x^{(i)})/\pi'_X(x^{(i)})$.

Then we proceed as before. Note that this generalises the strategy of the previous section, where the proposal distribution was taken to be $\pi_X(x)$. We only have to compute the likelihood and the two PDFs up to multiplicative constants, since the product of these will cancel out when we take the ratio of the weights.

How do we choose a good proposal distribution? One simple approach is to approximate the posterior distribution using numerical methods. Asymptotic theory suggests that as the amount of information in z^{obs} becomes large, so the prior becomes less and less important in determining the posterior, and the likelihood function tends to a Gaussian PDF (Schervish, 1995, sec. 7.4.2). Now this is unlikely to be true in the case of calibrating a hydrologic model: there is unlikely to be sufficient information in z^{obs}, particularly if we are realistic about the size of the model discrepancy. But the attraction of Importance Sampling is that the proposal distribution only needs to be approximately like the posterior. In fact, pragmatically, the proposal only needs to be a better approximation to the posterior than is the prior.

One simple approach is to take the proposal distribution to be a multivariate Gaussian distribution. The mean vector is the maximum likelihood value,

$$\hat{x} \triangleq \sup_{x \in \mathcal{X}} \log L(x), \tag{1.14}$$

and the variance matrix is the negative of the inverse of the Hessian matrix

of $\log L(x)$ evaluated at $x = \hat{x}$:

$$\hat{\Sigma} \triangleq -\left[\frac{\partial^2}{\partial x_i \, \partial x_j} \log L(x) \Big|_{x=\hat{x}} \right]^{-1}. \tag{1.15}$$

Both $\hat{x}$ and $\hat{\Sigma}$ can be assessed in a single numerical maximisation of the log-likelihood function: this maximisation does not have to be overly-precise. It is interesting to make the link back to more deterministic methods of model-calibration, in which finding $\hat{x}$, the best-fitting model-parameter, is seen as the goal. In this respect the Bayesian approach is clearly a generalisation: one that allows us also to assess the uncertainty in our choice of model-parameters (Rougier, 2005).

The Gaussian proposal is very tractable, and seems a safe choice because $\pi'_X(x) > 0$ for all $x \in \mathbb{R}^p$, so that the condition $\pi'_X(x) > 0$ wherever $\pi_X(x) > 0$ is automatically met. But there is a risk that the posterior distribution might have thicker tails than the proposal distribution, so that the condition that $L(x)\pi_X(x)/\pi'_X(x)$ is strictly bounded above might not be met. A simple and fairly robust expedient is to thicken the tails of the proposal distribution by switching from a multivariate Gaussian to a multivariate Student-t distribution with a small number of degrees of freedom (Geweke, 1989).

The multivariate Student-t is easy to sample from and easy to compute. Generally, if $X \sim \mathrm{N}_p(\mathbf{0}, I)$, $V \sim \chi^2(\delta)$, and $Y \sim \mathrm{T}_p(\mu, S, \delta)$, a p-dimensional multivariate Student-t where μ is the mean vector, S the scale matrix, and δ is the degrees of freedom, then

$$Y \stackrel{\mathcal{D}}{=} \mu + \frac{1}{\sqrt{V/\delta}} \, Q^T X \tag{1.16}$$

where Q is the Choleski decomposition of S, i.e. $S = Q^T Q$, and '$\stackrel{\mathcal{D}}{=}$' denotes 'equal in distribution'. In other words, the multivariate Student-t can be generated from p standard Gaussian variates and a single χ^2 variate. The variance of Y is $(\delta/(\delta-2))\, S$, and so we should choose δ and then set $S = ((\delta-2)/\delta)\, \hat{\Sigma}$ before computing Q, although setting $S = \hat{\Sigma}$ is conservative. The PDF of Y is

$$\pi_Y(y) \propto \left[1 + \delta^{-1}(z^T z)\right]^{-(\delta+p)/2} = \left[1 + v^{-1}(x^T x)\right]^{-(\delta+p)/2} \tag{1.17}$$

where $z \triangleq Q^{-T}(y - \mu) = \sqrt{\delta/v}\, x$. In other words, we can compute the PDF at y (up to a multiplicative constant, which we can ignore) using the sampled values x and v. Using an $n \times p$ matrix of X variates and an n-

vector of V variates, the whole set of sampled Y variates and PDF values can be generated in one simple operation.

Why stop at just one choice of proposal distribution? This procedure can be iterated, if we have a measure of how well our proposal distribution is matching the posterior distribution; this is known as *Adaptive Importance Sampling* (Oh and Berger, 1992). One simple measure is the *effective sample size*,

$$\text{ESS} \triangleq \frac{n}{1+\text{cv}^2} \quad \text{where} \quad \text{cv}^2 \triangleq \frac{\sum_{j=1}^{n}(w_j - \bar{w})^2}{(n-1)\bar{w}^2} \tag{1.18}$$

and $\bar{w}$ is the mean value of the weights. The ESS ranges from about 1, when a single $x^{(i)}$ dominates the weights, to n, when all weights are equal. It can be shown that the efficiency of the proposal distribution is roughly proportional to the ESS (Liu, 2001, sec. 2.5.3). Once we have a reasonably-sized sample from our initial proposal distribution, we can re-estimate the posterior mean and variance of X, and we can use these estimates to select a more appropriate proposal distribution, typically by updating the mean vector and scale matrix of the multivariate Student-t.

We can think of this approach as a pilot sample followed by the main sample, or we can iterate a few times, until the ESS has increased and stabilised at—one hopes—a value fairly close to n. We can use just the final sample to estimate properties of the posterior distribution or, if this is not sufficiently large, we can pool estimates from all the samples, weighting by the estimated standard error. We might also use the ESS to tune our choice of the degrees of freedom. Adaptive methods can sometimes be unstable, and so, if resources allow, a duplicate analysis would increase confidence in the result.

If after several iterations the ESS remains small, this suggests that our proposal distribution is a poor match to the posterior; i.e., the posterior is not unimodal and roughly bell-shaped. In this case a large n will be required, in order to raise the ESS to a reasonable value, or else a more sophisticated proposal can be used, such as a mixture of multivariate Student-t distributions (Oh and Berger, 1993). Another possibility is to transform one or more of the components of X. For example, if a component is strictly positive, then using a Gaussian marginal distribution for the logarithm might be better than, say, a Gamma distribution for the original value. Likewise, if a component is a proportion, then a Gaussian distribution for the logit might be better than a Beta distribution for the original value. Transforming X in this way, so that $\mathcal{X} = \mathbb{R}^p$ and $\pi_X(x)$ is relatively

symmetric, is is a good principle in general. Another one is to arrange, as far as possible, that the transformed values will be roughly uncorrelated in the posterior. Usually, this requires more information about the model than we possess, but we might be able to infer such a transformation from previous studies, if they have taken care to present multivariate uncertainty estimates for the model parameters.

The more advanced methods in this section are really concerned with making the most efficient use of a fixed budget of model-evaluations. The more efficient methods are a little more complicated to implement, and—taking a defensive view of the possibility of implementation errors—are only justified if the budget is quite constraining. Having said that, they can make a huge difference to the accuracy of the resulting approximations. The most robust and useful recommendation is to proceed in stages: spend some of the budget on a pilot sample from the prior distribution, and evaluate a diagnostic like the ESS. If this is a reasonable fraction of n, then it is quicker and safer to spend the rest of the budget on more points sampled from the prior; otherwise, extra efficiency can be purchased with extra coding.

1.5. Summary

The Bayesian approach provides a framework within which we may assess our uncertainties. It is important to appreciate that there is no unambiguous 'Bayesian answer' to a problem, and that the answer that we derive will be one that is imbued throughout by our judgements. This is obviously the case for the process of building the physical model. But it is also the case both of the formal process of describing and quantifying our beliefs about our physical model, the underlying system, and the observations; and of the calculations we implement to approximate features of our inferences regarding the model-parameters. To emphasise a point made at the start, once we have decided to assess our uncertainty probabilistically, we choose to adopt an approach such as estimating the posterior distribution $\pi_{X|Z}(x)$ because we think that it helps us to make better judgements about X, the model parameters, and it also helps us to convince other people to adopt these judgements, since our reasoning is transparent. In other words, calculating $\pi_{X|Z}(x)$ does not automatically lead us to the 'right' answer regarding X, but, rather, to a better answer than we might have got through other methods.

References

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001. Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association 96, 717–729.

Craig, P., Goldstein, M., Seheult, A., Smith, J., 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments. In: Gatsonis, C., Hodges, J., Kass, R., McCulloch, R., Rossi, P., Singpurwalla, N. (Eds.), Case Studies in Bayesian Statistics III. New York: Springer-Verlag, pp. 37–87, with discussion.

Davis, P., Rabinowitz, P., 1984. Methods of Numerical Integration, 2nd Edition. Orlando, Florida: Academic Press Inc.

Dawid, A., 1994. Foundations of probability. In: Grattan-Guinness, I. (Ed.), Companion Encyclopedia of the History and Philosophy of the Mathemetical Sciences. Vol. 2. London: Routledge, Ch. 10.16, pp. 1399–1406.

de Finetti, B., 1937. la prévision, ses lois logiques, ses sources subjectives. Annals de L'Institute Henri Poincaré 7, 1–68, see de Finetti (1964).

de Finetti, B., 1964. Foresight, its logical laws, its subjective sources. In: Kyburg, H., Smokler, H. (Eds.), Studies in Subjective Probability. New York: Wiley, pp. 93–158, English translation: H. Kyburg.

de Finetti, B., 1972. Probability, Induction and Statistics. London: Wiley.

DeGroot, M. H., Schervish, M., 2002. Probability and Statistics, 3rd Edition. Reading, Mass.: Addison-Wesley Publishing Co.

Evans, M., Swartz, T., 2000. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford: Oxford University Press.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57 (6), 1317–1339.

Gillies, D., 1994. Philosophies of probability. In: Grattan-Guinness, I. (Ed.), Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences. Vol. 2. London: Routledge, Ch. 10.17, pp. 1407–1414.

Goldstein, M., 2006. Subjective Bayesian analysis: Prinicples and practice. Bayesian Analysis 1 (3), 403–420.

Goldstein, M., Rougier, J., 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM Journal on Scientific Computing 26 (2), 467–487.

Goldstein, M., Rougier, J., 2006. Bayes linear calibrated prediction for complex systems. Journal of the American Statistical Association 101, 1132–1143.

Goldstein, M., Rougier, J., 2007. Reified Bayesian modelling and inference for physical systems. Journal of Statistical Planning and Inference, forthcoming as a discussion paper, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf`.

Jeffrey, R., 2004. Subjective Probability: The Real Thing. Cambridge University Press.

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B 63, 425–464, with discussion.

Kythe, P. K., Schäferkotter, M., 2004. Handbook of Computational Methods for Integration. Chapman & Hall / CRC.

Lad, F., 1996. Operational Subjective Statistical Methods. New York: John Wiley & Sons.

Liu, J., 2001. Monte Carlo Strategies in Scientific Computing. New York: Springer.

Oh, M.-S., Berger, J., 1992. Adaptive importance sampling in Monte Carlo integration. Journal of Statistical Computing and Simulation 41, 143–168.

Oh, M.-S., Berger, J., 1993. Integration of multimodal functions by Monte Carlo importance sampling. Journal of the American Statistical Association 88, 450–456.

Ripley, B., 1987. Stochastic Simulation. New York: John Wiley & Sons.

Robert, C., Casella, G., 1999. Monte Carlo Statistical Methods. New York: Springer.

Rougier, J., 2005. Probabilistic leak detection in pipelines using the mass imbalance approach. Journal of Hydraulic Research 43 (5).

Rougier, J., 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. Climatic Change 81, 247–264.

Savage, L., 1972. The Foundations of Statistics, revised Edition. New York: Dover.

Schervish, M., 1995. Theory of Statistics. New York: Springer, corrected

second printing, 1997.

Ziman, J., 2000. Real Science: What it is, and what it means. Cambridge University Press.