Journal of Microscopy, Vol. 190, Pts 1/2, April/May 1998, pp. 28–36.
Received 10 March 1997; accepted 24 June 1997

Probabilistic data analysis: an introductory guide

J. SKILLING
Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW, U.K.

Key words. Algorithm, Bayes, Gibbs, image reconstruction, inference, Metropolis, probability, simulated annealing, uncertainty.

Summary

Quantitative science requires the assessment of uncertainty, and this means that measurements and inferences should be described as probability distributions. This is done by building data into a probabilistic likelihood function which produces a posterior 'answer' by modulating a prior 'question'.

Probability calculus is the only way of doing this consistently, so that data can be included gradually or all at once while the answer remains the same. However, probability calculus is only a language; it does not restrict the questions one can ask by setting one's prior. We discuss how to set sensible priors, in particular for a large problem like image reconstruction.

We also introduce practical modern algorithms (Gibbs sampling, Metropolis algorithm, genetic algorithms, and simulated annealing) for computing probabilistic inference.

The problem

The deepening technology of science is systematically making us deal with uncertainty by taking the limitations of our measurements into account. This paper describes modern techniques for doing this.

Before even performing an observation, any experimenter is faced with equipment (such as an electron microscope). This equipment is designed to let the experimenter investigate some unknown features of observed samples. In electron microscopy, these unknowns $X = \{X_1, X_2, \ldots\}$ might represent the locations and strengths of the atomic columns in a thin crystal. Regrettably, equipment seldom measures X directly or fully. Instead, it produces data $D = \{D_1, D_2, \ldots\}$ which depend on X in some complicated way that might depend on awkward effects such as lens aberrations. Also, repeated observations even on the same sample usually give different data, owing to noise.

Our first requirement is to understand the behaviour of the equipment, so that we can model the passage

$$X \rightarrow D \qquad (1)$$

from unknown X to data D. Only then can we try to interpret a particular dataset D in terms of the unknown sample parameters X. Ideally, the data should be extracted as early as possible in the measuring process, to avoid any subsequent loss of information within the equipment. In practice, though, considerable internal processing is often carried out. This is harmless as long as the information relevant to X is preserved. For example, averaging several independent measurements of the same quantity is harmless. So is a Fourier transform, which is invertible without loss. If we write the equipment response in functional form as $D = \mathcal{R}(X)$, we might hope to invert via $X = \mathcal{R}^{-1}(D)$. Usually, though, the dataset will be incomplete: a range of different X would give the same data. This means that the passage from X to D is many-to-one, so that $\mathcal{R}^{-1}$ does not exist and the interpretation of the data in terms of X will not be unique.

Filters

Commonly, data are passed through some filter or other processing system $\mathcal{R}^*$ that is designed to behave as an approximate inverse of $\mathcal{R}$. Then $X^* = \mathcal{R}^*(D)$ is presented as 'the result'. One of the simplest filters is least squares, in which the result X^* is chosen to make the corresponding simulated or 'mock' data $\mathcal{R}(X^*)$ as close as possible to the actual data D, in the sense of least squares

$$X^* \text{ minimizes } \sum (\mathcal{R}_k(X) - D_k)^2. \qquad (2)$$

Weighted least squares includes weighting factors in the summation. Unfortunately, these attempts to follow the data as closely as possible interpret noise as signal. The consequent amplification of the effects of noise tends to have a catastrophic implication for the quality of least squares results.

More usefully, the data processing can be set up as an optimization, in which some quality functional $S(X)$ is maximized subject to some plausible figure-of-merit fit to the data: these techniques are called 'regularization' (Tikhonov, 1963; Titterington, 1985a, b). The commonest

functional is the sum of squares, $S(X) = -\Sigma X_i^2$. Allied to a square-law fit to the data, this functional yields a Wiener filter for calculating X^*. Another functional, particularly useful when X is the intensity of some distributed quantity, is the entropy (Gull & Daniell, 1979; Shore & Johnson, 1980; Livesey & Skilling, 1985), based on $S(X) = -\Sigma X_i \log X_i$. Quite apart from the sometimes arbitrary choice of functional, regularization requires setting a balance between the importance of quality S and of fitting the data. Various techniques, such as cross-validation (Stone, 1974), are proposed, but none is entirely convincing. The basic trouble is that any single 'result' will almost certainly be wrong in detail. Its quality and reliability may be adequate for limited purposes, yet it will be possible to glean more from the data.

Probability

In quantitative treatment, we should acknowledge the uncertainty introduced by noise. Even with fixed inputs X, the corresponding data would vary according to some probability distribution

$$\Pr(D|X) = \text{likelihood function} \tag{3}$$

where '|' means 'conditional upon', or 'supposing' for short. Usually, we suppose that noise is additive, so that

$$D = \mathcal{R}(X) \pm \sigma. \tag{4}$$

Often, additive noise is taken to be Gaussian, in which case we have the explicit formula

$$\Pr(D|X) = Z^{-1} \exp(-\chi^2/2) \tag{5}$$

where

$$\chi^2(X) = \sum_k (\mathcal{R}_k(X) - D_k)^2/\sigma_k^2 \tag{6}$$

is the usual chi-squared misfit between mock data and actual data, scaled into a probability by the standard deviations σ, and where $Z = \Pi(2\pi\sigma_k^2)^{1/2}$ is the overall normalization.

The likelihood function is the formal specification of the behaviour of the observing equipment, usually given as a responsivity $\mathcal{R}$ allied to noise σ. It is needed for a sufficient range of inputs X to cover all admissible interpretations of data D that we may observe. Ideally, it would be assigned by repeated direct measurements over the (huge) range of allowable inputs. In practice, $\mathcal{R}$ is usually assigned from a theoretical description of the equipment's performance, appropriately calibrated by reference to a realistically small number of known inputs. Sometimes, the equipment cannot be reliably described in full, in which case the description may include some *ad hoc* extra factors X'. We absorb these into the set of unknown inputs X. One fairly common case is the one of poorly known noise level, where an appropriate scaling factor needs to be used in the likelihood function as an extra unknown. An unpredictable contrast level might be another such parameter. The presence of such factors in the data certainly degrades the quality of inferences that might otherwise be made. However, omitting them from the analysis is even more damaging because that tends to give results with systematic errors and spuriously high supposed accuracy.

Once we have the likelihood function, we know what the data mean, so can realistically aim to infer the inputs X. Such inference will be uncertain, so we can at best aim for the posterior (i.e. after the data) probability distribution

$$\Pr(X|D) = \text{inference}. \tag{7}$$

This involves inverting the assigned likelihood $\Pr(D|X)$. But, as we see below, inversion requires us to assign a *prior* probability distribution

$$\Pr(X) = \text{prior}. \tag{8}$$

The prior represents our guess, or pre-conception, about the range of unknown values that might be present, before we use the data. It is an interesting philosophical fact that we need this preliminary hypothesis about what we are observing before we can make sense of our observations: the Universe does not give us absolute results. Prior information can be very relevant, and is always necessary.

Bayes' theorem

We now use the ordinary rules of probability calculus. The joint probability

$$\Pr(X, D) = \Pr(X)\Pr(D|X) = \Pr(D)\Pr(X|D) \tag{9}$$

is the essential distribution from which all else follows. It is summed (integrated) to obtain the prior predictive value

$$\Pr(D) = \sum_X \Pr(X)\Pr(D|X) = \text{evidence} \tag{10}$$

which is a single number having dimension $[D^{-1}]$. Dividing by this, we reach 'Bayes' theorem' for the inference

$$\begin{array}{ccc} \Pr(X|D) = \Pr(X)\Pr(D|X)/\Pr(D) & \propto & \Pr(X)\Pr(D|X). \\ \text{inference} & \propto & \text{prior} \times \text{likelihood} \end{array} \tag{11}$$

(Experts in probability will recognize that we have omitted the background information I on which all these probabilities should be conditioned.) In short:

1 Ask a question by setting a prior $\Pr(X)$.
2 Describe your equipment by its likelihood function $\Pr(D|X)$.
3 Acquire your particular data D.
4 Optionally, calculate the evidence $\Pr(D)$ as a measure of the quality of your question.
5 Evaluate your answer as the posterior inference $\Pr(X|D)$.

Probability calculus is unique; it is the only calculus

within which uncertainties about propositions are manipulated consistently with the logical (TRUE/FALSE) status of the propositions. In particular, any posterior inference can be used as a prior when analysing extra data, so that the order in which one uses the data is immaterial (as it should be). Kolmogorov (1950) is widely quoted as the author of the axiomatic basis of probability calculus, but it was Cox (1946, 1961) who showed that no other calculus is admissible. The only freedom is to take some monotonic function instead, such as $100\,\Pr(\bullet)$ (percentage) or $\Pr(\bullet)/(1-\Pr(\bullet))$ (odds), but such changes are merely cosmetic. It follows that other methods are either equivalent to probability calculus (in which case they are unnecessary), or are wrong. Thus in fuzzy sets, an assignment of fuzzy membership is equivalent to assigning a probability of membership, so is harmless. However, insofar as the various rules of fuzzy logic (Zadeh, 1965; Klir & Folger, 1988) diverge from probability calculus, logical error enters and practical power is lost. In particular, a fuzzy deduction from part of the data would not give a consistent starting point for using the remainder of the data.

Interestingly, probability calculus never instructs us to maximize or optimize any unknown parameter. Instead, the rules instruct us to sum or integrate over unknown quantities, so that their effect is averaged over all plausible values. If we think of the inference $\Pr(X|D)$ as defining the shape of a mountain, then our task is to find and explore the summit area of reasonably plausible X, not just to locate the single highest point of maximum probability. After all, it is easy to envisage asymmetrical probability distributions in which the highest point is a narrow peak far from the bulk of the distribution. Maximization may be convenient because only one value need be considered thereafter, and it may also be fairly harmless if the parameter in question is quite closely constrained – but it is always an approximation to be handled with care. Direct use of probability calculus is often called the 'Bayesian' method. If the analysis includes some maximization or similar approximation, the methodology is called 'empirical Bayesian'.

Prior probabilities

The language of probability calculus is defined, but there is no restriction on the questions one can ask by posing a suitable prior. This freedom, nevertheless, can be assessed objectively through the evidence value. Different priors will induce different evidence values $\Pr(D)$ (strictly, $\Pr(D|$prior and other assumptions)). An inappropriate prior shows up through a relatively low evidence value. The art of setting a prior is to allow sufficient flexibility to cover all the inputs X that one might plausibly need in order to fit the data reasonably, without widening the field so much that the particular data D will become intrinsically implausible. In other words, ask a question appropriate to your data.

In practice, the choice of prior is often guided by symmetry considerations, intrinsically equivalent X being assigned the same prior values. The standard introductory problem has an ordinary cubical 'fair' die that is thrown so that any of its faces ($X=1,2,\ldots,6$) may fall upwards. There are only six states, so the prior consists of just six numbers $\Pr(X_1),\ldots,\Pr(X_6)$. These numbers sum to 1 because it is supposed to be certain that one of the results actually occurs. The natural assignment $\Pr(X_1)=\ldots=\Pr(X_6)=1/6$ is justified by the symmetry of the problem. Other than marking, there is no physical difference between the faces, so our pre-conception about the states is invariant with respect to interchange. Unless we use the obvious uniform assignment, this invariance would be broken.

A less precise example concerns the temperature of water under standard conditions. Here the temperature X can take any value between 0 °C and 100 °C. Although there is no justifying symmetry between different temperatures, it might still be reasonable to take a uniform flat prior $\Pr(X)=1/100$ in $0\le X\le 100$. Yet this is guided by convention as much as by physical reality. If we were more used to thinking of temperature in terms of absolute coolness, $X'=1/(X+273{\cdot}16)$, then we would more naturally set $\Pr(X')$ constant within $1/373{\cdot}16\le X'\le 273{\cdot}16$. Transforming back to the Celsius scale by $\Pr(X)\,\mathrm{d}X=\Pr(X')\,\mathrm{d}X'$, we would find that $\Pr(X)$ was no longer constant. This example warns that a flat prior is not necessarily 'correct'. Indeed, depending on circumstances, we might well expect some temperatures to be more likely than others; if so, we are entitled to encode this into our prior. Jaynes (1968) persuasively develops a similar example.

Another example concerns the brightness X of a light source. Here X must be positive, and physics defines a natural scale (that of power output, in which brightnesses are additive), but there is no obviously useful upper limit to X. Loosely, one might try to assign a flat prior, $\Pr(X)$ constant in $0\le X<\infty$, and hope to avoid the awkward corollary that normalization requires the constant to be zero. After all, as soon as one measures realistic data about X, the posterior 'answer' $\Pr(X|D)$ should become well-behaved. However, the prior predictive evidence number, $\Pr(D)$, will still be zero, reflecting the fact that X has initially zero probability of lying in any finite range. So there is no real escape from assigning a properly normalized prior, appropriate to the situation. Here, we cannot set a proper prior until we assign at least the overall expected magnitude of X. Anybody who tried to use the improper flat prior would lose by an infinite factor in the objective comparison with the evidence of a competitor who acknowledged a vague idea of the range of brightnesses being considered. We have to ask a sensible question before we can get a sensible answer.

The Principle of Maximum Entropy (Cox, 1961; Jaynes,

1978) is a sophisticated extension of a symmetry argument, applicable where some average property or properties of X are to be controlled. According to the principle, $\Pr(X)$ may be assigned by maximizing its entropy subject to any 'ensemble average' constraints that may be available:

$$\text{Maximize } S = -\int \Pr(X) \log \Pr(X) \mathrm{d}X$$

$$\text{subject to} \quad \int c_k(X) \Pr(X) \mathrm{d}X = C_k,\ k = 1, 2, \ldots . \qquad (12)$$

For example, suppose that the light sources we were considering above might have some average brightness B. This general guidance can be captured as an average constraint $\int X \Pr(X)\, \mathrm{d}X = B$ which (together with normalization) yields the prior $\Pr(X) = B^{-1} \exp(-X/B)$. Many other forms might have been suggested (Cauchy, Gaussian, ...), but maximum entropy gives an objective reason for preferring the exponential form.

Image reconstruction

Image reconstruction is a more difficult example, in which there is a separate brightness X_i for each pixel i of the image. There may well be correlations, whereby neighbouring pixels are expected to have similar brightnesses so that the image is expected to be locally smooth; Ripley (1988) explores such possibilities. On the other hand, if the image may have sharply localizable sources, it can be better to ignore correlations and treat each pixel as independent, by writing $\Pr(X) = \Pi\, p(X_i)$ where $p(\bullet)$ is the prior we need to assign to the brightness of a single pixel. It is tempting to assign an expected average brightness b to a pixel, and appeal to the principle of maximum entropy to write $p(x) \propto \exp(-x/b)$. An even simpler suggestion is to let $p(x)$ be constant up to some assigned maximum brightness, but these prescriptions fail.

In passing to image reconstruction, we have reached a problem in which there may well be more unknowns (pixel values) than there are reliable data. Indeed, we wish to able to use a very large number of arbitrarily small pixels so that the reconstruction is not visibly blocked. This means that the reconstruction problem is heavily under-constrained. The data can constrain at best a minority of the unknowns, so most of the reconstruction must be controlled by the prior. The choice of prior is no longer just an esoteric detail; it can dominate the results. If we used the exponential prior $\exp(-x/b)$ on individual independent pixels, we would find that the overall brightness on a macroscopic domain of N pixels would converge around Nb, with a plausible degree of variability that collapsed towards zero (relatively as $\mathcal{O}(N^{-1/2})$) as the pixel size was reduced and N thereby increased. In other words, simply by deciding to compute with smaller pixels, we would become increasingly convinced that the macroscopic brightness pattern was just flat and uniform. It is a consequence of the law of large numbers, which can only be evaded by ensuring that only a limited number of pixels have appreciable brightness.

It is possible to set priors that make sense, in that macroscopic structure can appear regardless of subsidiary pixellation. Such a prior is called a 'process' (Kingman, 1993). The prior $p(x|h)$ must depend not only on the brightness x but also on the pixel size h in such a way that subsidiary pixellation is immaterial. The process most commonly suggested is the Gamma process, essentially $p(x|h) \propto x^{1+h} \mathrm{e}^{-x}$) (e.g. Sibisi & Skilling, 1997). If the pixel size is allowed to shrink, this prior concentrates on small brightnesses x in just such a way that only a limited number of pixels have appreciable brightness. Typical brightness patterns are 'atomic'. Of course, this does not mean that any mean reconstruction, averaged over all plausible brightness patterns according to the posterior $\Pr(X|D)$ would be sharp and atomic. To the contrary, the mean reconstruction will be only as sharp as the data demand.

Another process, which is completely atomic and hence easier to compute, has

$$p(x|h) = (1 - h)\delta(x) + h \exp(-x), \quad (h \text{ small}) \qquad (13)$$

after scaling h and x to take out their dimensional units A and B. Most small pixels have zero brightness because of the Dirac delta function. The others have brightnesses distributed exponentially as $\exp(-x/B)$, that being the natural maximum entropy form. Pixels with finite brightness appear randomly with probability h/A in small size h, so that they have a Poisson distribution in location. Because the brightness is entirely contained in a finite set of 'point mass' delta functions, this form is being called the 'massive inference' prior.

Example

To illustrate the use of probability theory, let the unknown X be simply the temperature of a liquid, which might be water or ethanol.

1 Ask a question by setting a prior $\Pr(X)$. Suppose that the liquid is water. As before, this assumption naturally leads to the prior $\Pr(X) = 0{\cdot}01$ in $0 \le X \le 100$, zero otherwise.

2 Describe your equipment by its likelihood function $\Pr(D|X)$. The equipment is a thermometer whose readings D may differ from the temperature X by up to 5 °C. Hence the likelihood is $\Pr(D|X) = 0{\cdot}1$ in $X - 5 \le D \le X + 5$, zero otherwise. Realistic equipment more commonly gives roughly Gaussian errors, but we take a uniform distribution for introductory simplicity. Note the normalization $\int \Pr(D|X)\, \mathrm{d}D = 1$ common to all probabilities.

3 Acquire your particular data D. The thermometer records $D = -3$ °C.

4 Optionally, calculate the evidence from Eq. (10):

$\Pr(D) = 0{\cdot}002\,(^\circ\mathrm{C})^{-1}$ is a measure of the quality of your question. Technically, we should write the evidence as $\Pr(D|\text{water})$ because we have assumed that the liquid is water.

5 Evaluate your answer as the posterior inference from Eq. (11): $\Pr(X|D) = 0{\cdot}5$ in $0 \le X \le 2$, zero otherwise. This answer is sensible, because X is known from the prior to be greater than 0, and from the data to be less than 2. Technically, we should write the inference as $\Pr(X|D, \text{water})$.

Now suppose instead that the liquid is ethanol, for which the prior range is $-80 \le X \le 80\,^\circ\mathrm{C}$. Repeating the analysis yields a larger evidence $\Pr(D|\text{ethanol}) = 0{\cdot}00625$, with a different inference $\Pr(X|D) = 0{\cdot}1$ in $-8 \le X \le 2$, zero otherwise. Clearly the thermometer reading $(-3 \pm 5)\,^\circ\mathrm{C}$ suggests ethanol rather than water, but the Bayesian analysis quantifies the relative preference. Indeed an evidence value $\Pr(D|\text{assumptions})$ is nothing more than a likelihood value $\Pr(D|\bullet)$ for the current assumptions. If we wish to compare different assumptions (such as water versus ethanol), we just assign prior probabilities to these assumptions and repeat the probabilistic analysis at this higher level. The same calculus still applies.

Computation

The logical necessity of using Bayes' theorem for inference is inescapable, yet until quite recently probability calculus has not been widely used for large problems. This is partly because of a primitive desire for absolute results, leading to a reluctance to acknowledge the role of the prior. It is also because the probability approach forces us to contemplate exploration of all reasonably plausible X. Yet whenever X has more than a very few components, full exploration of the huge space of $\Pr(X|D)$ is wildly impractical: the apparently exponential cost of introducing more components is sometimes called 'the curse of dimensionality'.

Markov chain Monte Carlo

The trick is to use Monte Carlo methods (Hammersley & Handscomb, 1964). Suppose that several (n) samples $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ can be found, drawn independently from the posterior $\Pr(X|D)$. Surprisingly, this extremely impoverished selection from the huge space is often sufficient. In practice, we are nearly always interested primarily in simple numerical properties $f(X)$ such as an average separation between components, or a mean intensity ratio, or perhaps an individual component like X_5. The n samples give us n independent estimates $f(X^{(1)}), f(X^{(2)}), \ldots, f(X^{(n)})$ of this property. These n numbers allow us to estimate the basic statistics of f, such as its mean μ and standard deviation σ uncertainty. The mean

$$\langle f \rangle = n^{-1} \sum_{i=1}^{n} f(X^{(i)}) \tag{14}$$

estimates μ with an uncertainty of about $\sigma/\sqrt{n}$ (assuming Gaussian statistics), while the variance

$$\mathrm{var}(f) = \langle (f - \langle f \rangle)^2 \rangle \tag{15}$$

estimates σ^2. Although Monte Carlo results only converge very slowly, as $\mathcal{O}(n^{-1/2})$, to their precise asymptotic limits, the major estimates soon settle down to within their error bars to high probability. For example, with $n = 16$ independent samples, the mean $\langle f \rangle$ estimates μ with standard deviation around $\sigma/4$, so that only once in about 15 000 trials should the estimate be in error by more than σ. This is adequate, because if μ represents a quantity that is (by definition) uncertain by σ, it need not be calculated to any greater precision. Non-Gaussian statistics generally give rather slower convergence, and error bars are estimated less accurately than primary quantities, but it remains true that Monte Carlo methods can give the inferences we seek. The computations are ambitious, but feasible. There is an inherent limitation of Monte Carlo methods, in that full certainty is never achieved no matter how many samples are computed. The probability of serious algorithm error, though, decreases exponentially with the number of samples, so that the calculation can be made more reliable than the data being used.

Computing a random sample X from a complicated joint distribution

$$\pi(X) \equiv \Pr(X, D) \propto \Pr(X|D) \tag{16}$$

is usually impossible to do directly. The plausible X will be too small a fraction of the possible X to find by immediate exploration. Instead, algorithms operate incrementally, holding a single sample and evolving it 'randomly' (with a pseudorandom generator) so that its probability distribution approaches the desired $\pi(X)$. A sequence in which the next state depends partly on the current state is a Markov chain, so such algorithms are called Markov chain Monte Carlo (MCMC) methods (Hastings, 1970; Gelfand & Smith, 1990). As a formal detail, any MCMC algorithm should be capable of reaching all states (or at least those within the support of π); all recommended algorithms can do this. Smith (1991), Smith & Roberts (1993) and Besag & Green (1993) are good sources.

There is an analogy with statistical mechanics, where a physical system evolves under essentially random influences towards a thermal equilibrium described by

$$\pi(X) \propto \exp(-\lambda E(X)) \tag{17}$$

where E is energy and λ is reciprocal temperature, conventionally set to 1. If π is thought of as a mountain to be climbed, then E is a corresponding basin to fall into.

The algorithm will start with some state $X^{(0)}$ which will usually be highly atypical of π; likewise a physical system may be released far from equilibrium. The physical system will relax, and similarly the algorithm's state will lose memory of any special initial conditions and 'burn in' towards equilibrium. Thereafter, the physical system will have reached thermal equilibrium and will ergodically explore the states available to it according to the distribution π. The algorithm will also explore its distribution so that each sample X is an unbiased, random sample from $\pi(X)$.

MCMC algorithms all have a common structure. A state X induces a potential successor Y through some pseudo-random transition scheme

$$X \to Y \quad \text{according to the probability} \quad T(Y|X) \equiv \Pr(Y|X) \tag{18}$$

that defines the algorithm being used. Had the system been in state Y to start with, it would be able to move to other states, in particular reverting to X with

$$Y \to X \quad \text{according to the same probability law} \quad T(X|Y). \tag{19}$$

Taken together, these transitions would favour an equilibrium in which X and Y appear with relative frequency

$$\frac{\Pr(Y)}{\Pr(X)} = \frac{T(Y|X)}{T(X|Y)} \tag{20}$$

because the numbers of forward and backward transitions then balance. Actually, we seek a procedure in which $\Pr(X)$ mimics the desired distribution $\pi(X)$, so we want $\Pr(Y)/\Pr(X) = \pi(Y)/\pi(X)$ and the suggested transition scheme T may not have that property.

To correct the algorithm, transitions are performed only with an an appropriate acceptance probability $\alpha(Y|X)$, chosen to ensure that

$$\frac{\Pr(Y)}{\Pr(X)} = \frac{T(Y|X)\alpha(Y|X)}{T(X|Y)\alpha(X|Y)} = \frac{\pi(Y)}{\pi(X)}. \tag{21}$$

The acceptance probabilities should be as large as possible, to avoid the waste of computing potential transitions that are rejected. However, no probability can exceed 1. Hence we set

$$\alpha(Y|X) = \min\left(1, \frac{T(X|Y)\pi(Y)}{T(Y|X)\pi(X)}\right) \tag{22}$$

which has the desired effect with minimal waste. All transitions, between any X and Y, relax towards the desired equilibrium; this property is known as 'detailed balance'.

Gibbs sampling

Gibbs sampling, named by its authors Geman & Geman (1984) after the physicist J. W. Gibbs, can be the easiest to implement. In its basic form, it changes just one component of X, say the ith component X_i, at a time. All other components preserve their values ($Y_j = X_j$ for $j \neq i$) but the ith component is chosen by re-sampling from the marginal distribution

$$\begin{aligned} T(Y_i|X) &= \Pr(Y_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots) \\ &\propto \pi(X_1, \ldots, X_{i-1}, Y_i, X_{i+1}, \ldots) = \pi(Y). \end{aligned} \tag{23}$$

With this algorithm, the acceptance probability is always 1, so transitions are never rejected. The component i can either cycle round ($i = 1, 2, 3, \ldots, 1, 2, 3, \ldots$) or be selected at random. Gibbs sampling presupposes that the marginal distribution has some accessible algebraic form to sample from. If this has to be approximated, perhaps by interpolating between several potential values of Y_i, then the acceptance probability should be re-introduced to correct any resulting imbalance.

Gibbs sampling yields a particularly slow algorithm if there is strong coupling between components of X. For example, suppose the sum $X_1 + X_2$ is tightly constrained through an accurate measurement. The initializing guess $X^{(0)}$ will usually be highly atypical of the desired distribution, so that X_1 and X_2 will be far from equilibrium. In the first Gibbs step, X_2 will keep X_1 oppositely far from equilibrium, then X_1 will keep X_2 away, and so on for many steps. Generally, Gibbs algorithms are slow whenever the 'mountain' $\pi(X)$ is high, and has sharp ridges that are not conveniently aligned with the coordinate axes.

Metropolis algorithm

The random-walk Metropolis algorithm (Metropolis *et al.*, 1953) avoids the Gibbs dependence on a particular coordinate system, and is also easy to implement. Here X is incremented by a random vector of arbitrary direction and some scale s (usually Gaussian in each coordinate):

$$Y_i = X_i + r_i, \quad r_i \in \mathcal{N}(0, s^2). \tag{24}$$

This transition scheme is symmetric, $T(Y|X) = T(X|Y)$, so that the acceptance probability simplifies to

$$\alpha(Y|X) = \min(1,\ \pi(Y)/\pi(X)). \tag{25}$$

If s is given too small a value, nearly every step will be accepted (because Y will be close to X), but the random walk will need many small steps before samples become effectively independent. On the other hand, if s is given too large a value, then most potential transitions will damage the probability value ($\pi(Y) \ll \pi(X)$) and thereby be rejected. Some compromize value should be used, perhaps aiming for an acceptance probability $\alpha \sim 1/3$.

Generally, Metropolis algorithms are most efficient when the step size allows the probability ratios to be $\mathcal{O}(1)$. In terms of the energy analogy, $\Delta E \sim 1$. Yet the energy loss involved in reaching the floor of the basin may be much

greater than 1, especially in large problems with many data. To fit 1000 data each accurate to 1% ($=e^{-4\cdot6}$) involves an energy change of 4600. Hence thousands of Metropolis steps seem to be needed for the burn-in alone, even before the energy basin is explored.

Genetic algorithms

Genetic algorithms (Davis, 1991) treat two or more samples $X^A, X^B, \ldots$ as a single algorithm state. Potential transitions include exchanges of one or more coordinate values between the samples, as for example

$$\begin{aligned} &\{X^A = (a,b,c),\ X^B = (d,e,f)\} \\ \rightarrow\quad &\{Y^A = (d,b,c),\ Y^B = (a,e,f)\} \end{aligned} \tag{26}$$

where the first (or any other) coordinate(s) out of three are exchanged. Genetic algorithms tend to be recommended for problems in which the unknowns X_i are discrete rather than continuous, and in which there may be distinct and separated summits of plausible results. Once again, though, the algorithm is slow if the probability factors are large, because substantial steps are then usually rejected.

Convergence

Gibbs, Metropolis and genetic algorithms are the common generic forms of MCMC. Yet any particular problem may have its own particular structure that might be used to design a transition scheme that allowed larger steps, and was hence faster. In the example above where the sum of two components, $X_1 + X_2$, was measured accurately, a transition with X_1 and X_2 changing equally and oppositely might well be advantageous. MCMC algorithms can also be mixed by alternating steps from different procedures, so that if one procedure becomes inefficient another might be able to circumvent the problem. Unfortunately, there is as yet no useful theory on how best to design an efficient MCMC algorithm; it is a matter for heuristic ingenuity (Gelman *et al.*, 1995). The essential requirements are that the overall procedure is capable of reaching all relevant states X, that detailed balance is achieved by appropriate rejection and that 'enough' steps are taken.

The algorithm should be continued past burn-in until some reasonable number of independent samples ought to have been obtained. Because a state retains memory of its predecessors, many intervening steps may be needed before samples become effectively independent. Progress has been made (Roberts, 1992), but unfortunately there are as yet no generally useful formal estimates of convergence times. Thus, in practice, MCMC algorithms are continued until the spread of samples appears to have converged, according to some *ad hoc* criterion. One such criterion is that the log-likelihood $\log \Pr(D|X)$ fit to the data no longer appears to drift. However, apparent convergence might be spurious. The danger is that if an algorithm is not given sufficient time, it may appear to have converged, but to the wrong result and quite likely with apparently high but spurious precision. Some authors (Gelman & Rubin, 1992a, b) recommend that MCMC results be confirmed by running the algorithm with several different starting positions $X^{(0)}$, and checking that the desired inferences $f(X)$ are consistent; if not, the computation time was too short. Such checking can reveal failure, though consistency does not formally guarantee convergence.

Simulated annealing

In nearly all large problems, the π mountain is high, as well as having an awkward shape. Equivalently, the energy basin is deep. In these circumstances, MCMC algorithms tend to be slow because they tend to use $\mathcal{O}(1)$ energy changes. Simulated annealing (Kirkpatrick *et al.*, 1983; Aarts & Kost, 1989) is a general way of ameliorating this. Instead of working directly with π (or the energy E at unit temperature), work at some higher temperature using $\lambda < 1$. The correct distribution π is replaced by π^λ. The idea is that this flattens the mountain, so that the algorithm can reach the higher parts more quickly. Final exploration needs λ to have its correct value of 1, but the burn-in time spent reaching the summit area can be much reduced by bringing λ up to 1 gradually, according to some 'cooling strategy'.

Technically, it is better to keep the prior intact and anneal the likelihood function alone, giving

$$\pi_\lambda(X) \equiv \Pr(X)\Pr(D|X)^\lambda, \qquad 0 \le \lambda \le 1. \tag{27}$$

Flattening the prior as well leads to the distribution π^λ, which would be impossible to sample from if its integral became divergent. Flattening just the likelihood corresponds to bringing the data in gradually, from absence ($\lambda = 0$) to full effect ($\lambda = 1$). It also suggests using a random sample from the prior itself as a natural starting sample $X^{(0)}$, since this is correctly unbiased at the start when $\lambda = 0$.

At every stage the distribution π_λ is integrable, and can be used to define

$$\langle \log \Pr(D|X) \rangle_\lambda = \frac{\int \log \Pr(D|X)\, \pi_\lambda(X)\, \mathrm{d}X}{\int \pi_\lambda(X)\, \mathrm{d}X}. \tag{28}$$

Just like any other function $f(X)$, this average is estimated by summing over samples obtained at reciprocal temperature λ. These samples may already have found a use in trying to check convergence at λ, but there is another unexpected use from the identity

$$\begin{aligned} \Pr(D) &= \int \Pr(X)\Pr(D|X)\,\mathrm{d}X \\ &= \exp \int_0^1 \langle \log \Pr(D|X) \rangle_\lambda \,\mathrm{d}\lambda = \text{evidence}. \end{aligned} \tag{29}$$

Simulated annealing thus allows the Bayesian analysis to be completed by calculating the prior predictive 'evidence' number that enables comparison with other analyses of the same data using different assumptions.

Multi-modality

Suppose the π-mountain has not just one but two separated summits (or even more), corresponding to an ambiguity between distinctly different results X. Fourier data with unknown phases, as in electron diffraction and X-ray crystallography, are particularly prone to such ambiguity. Without simulated annealing, a sample started at the base of the mountain will tend to move uphill, so that the chance of ultimately finding it around a specific summit will be roughly given by the size of that summit's footprint (equivalently, the size of the energy basin of attraction).

With simulated annealing, the mountain is initially flat and the samples are constrained by the prior alone. As the temperature falls, the mountain grows and the samples become effectively confined to the higher parts. Below some separation temperature, samples become largely confined above the pass between the summits. Beyond separation, there is the technical possibility of tunnelling between summits if the transition scheme allows, but in practice the chance of tunnelling tends to switch off exponentially. At separation, the chance that a sample is located around a specific summit is, therefore, roughly determined by the size of that summit above the separation contour at the height of the pass. Even this is not the right answer.

Actually, we want the sampling density to reflect π itself, with the chance that a sample is located around a specific summit being the evidence integral over that summit – a product of height π and volume ΔX in which either factor could dominate. Unless the MCMC algorithm allows communication between the summits, the only recourse seems to be to use the evidence integrals accumulated along runs from different starting points in order to assess the relative importance of the different summits. This may be tedious, and is not usually attempted, but the formalism allows it.

Conclusions

Probability calculus is a major unifying principle in the description of uncertainty. It is the only consistent calculus, and requires us to cast all uncertainty into the standard form of a probability distribution. In data analysis, this uncertainty includes our original uncertainty about the unknown parameter(s) X that we seek. We have to cast this as a prior 'question' $\Pr(X)$. In small problems where X is a simple parameter such as a temperature, details of the prior may well influence the posterior 'answer' $\Pr(X|\,\text{data})$ only slightly. In large problems such as image reconstruction, an incorrectly chosen prior can be disastrous. The principle remains that one should try to ask a question appropriate to one's data, but application of the principle can require care.

Full probabilistic analysis is an ideal that is not always attained in practice. For a start, it may be too difficult to model the behaviour of the equipment accurately, even with the aid of auxiliary unknown variables. One may then be forced to use some figure of merit instead of the desired likelihood function. Much of the power of probability calculus is thereby lost, but computations might still produce a useful result.

Markov chain Monte Carlo (MCMC) algorithms form the basis of practical computation of probability distributions in several or many unknowns. MCMC results are given as a sequence of at least several typical samples $X^{(1)}, X^{(2)}, \ldots$ drawn from the posterior distribution. If a single X is required as 'the' result, then the mean $\langle X \rangle$, obtained by averaging the samples, is usually recommended. The single most probable X, obtained by maximizing the posterior, can easily be atypical of the distribution as a whole, so that its properties $f(X)$ are prone to bias. The mean is generally more robust and useful than the maximum.

The basic MCMC methods are easy to program and implement. No complicated gradient or maximization procedure is involved – just the local evaluation of prior and likelihood probability values. The lack of a formal convergence criterion is usually just a technical quibble that can be ameliorated by running the algorithm for longer, or from different starting points, and checking the major results for consistency. Finally, the simulated annealing variant of MCMC is capable of calculating the prior predictive 'evidence' value, should that be required.

References

Aarts, E.H.L. & Kost, J. (1989) *Simulated Annealing and Boltzmann Machines*. Wiley, New York.

Besag, J. & Green, P.J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B.* **55,** 25–37.

Cox, R.T. (1946) Probability, frequency, and reasonable expectation. *Am. J. Phys.* **14,** 1–13.

Cox, R.T. (1961) *The Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore, Maryland.

Davis, L. (1991) *Handbook of Genetic Algorithms*. van Nostrand Reinhold, New York.

Gelfand, A.G. & Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85,** 398–409.

Gelman, A., Roberts, G. & Gilks, W. (1995) Efficient Metropolis jumping rules. *Bayesian Statistics* (ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith) Vol. 5. Oxford University Press.

Gelman, A. & Rubin, D.B. (1992a) A single series from the Gibbs sampler gives a false sense of security. *Bayesian Statistics* (ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith) Vol. 4. Oxford University Press.

Gelman, A. & Rubin, D.B. (1992b) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7,** 457–472.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **6,** 721–741.

Gull, S.F. & Daniell, G.J. (1979) The maximum entropy method. *Image Formation from Coherence Functions in Astronomy* (ed. by C. van Schooneveld), pp. 219–225. Reidel, Dordrecht.

Hammersley, J.M. & Handscomb, D.C. (1964) *Monte Carlo Methods*. Methuen, London.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their application. *Biometrika,* **57,** 97–109.

Jaynes, E.T. (1968) Prior probabilities. *IEEE Trans. Systems Sci. Cybernet.* **SSC-4,** 227–241.

Jaynes, E.T. (1978) Where do we stand on maximum entropy? *Reprinted in E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics* (ed. by R. Rosenkrantz), 1983, pp. 211–314. Reidel, Dordrecht.

Kingman, J.F.C. (1993) *Poisson Processes*. Oxford University Press.

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983) Optimization by simulated annealing. *Science,* **220,** 671–680.

Klir, G. & Folger, T. (1988) *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, New York.

Kolmogorov, A.N. (1950) *Foundations of the Theory of Probability*. Chelsea Publishing Co., New York.

Livesey, A.K. & Skilling, J. (1985) Maximum entropy theory. *Acta Crystallogr.* **A41,** 113–122.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machine. *J. Chem. Phys.* **21,** 1087–1091.

Ripley, B.D. (1988) *Statistical Inference for Spatial Processes*. Cambridge University Press.

Roberts, G.O. (1992) Convergence diagnostics of the Gibbs sampler. *Bayesian Statistics* (ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith), Vol. 4, Oxford University Press.

Shore, J.E. & Johnson, R.W. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minumum cross-entropy. *IEEE Trans. Info. Theory,* **IT-26,** 26–39; **IT-29,** 942–943.

Sibisi, S. & Skilling, J. (1997) Prior distributions on measure space. *J. R. Statist. Soc.* **59,** 217–235.

Smith, A.F.M. (1991) Bayesian computational methods. *Philos. Trans. R. Soc. London A,* **337,** 369–386.

Smith, A.F.M. & Roberts, G.O. (1993) Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B,* **55,** 3–23.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc. B,* **36,** 111–147.

Tikhonov, A. (1963) Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.* **4,** 1035–1038.

Titterington, D.M. (1985a) Common structure of smoothing techniques in statistics. *Int. Statist. Rev.* **53,** 141–170.

Titterington, D.M. (1985b) General structure of regularization procedures in image processing. *Astron. Astrophys.* **144,** 381–387.

Zadeh, L.A. (1965) Fuzzy sets. *Information and Control,* **8,** 338–353.