Appendix: Bayes-MCMC Inference

Here we present a primer on the Bayesian paradigm for statistical inference coupled with MCMC algorithms to implement such techniques.

A.1 Bayes Theorem

Say that observable quantities X are thought to be distributed according to a density function $f(x|\theta)$, where θ is an unknown parameter vector. Of course, such use of probability distributions for data given parameters is ubiquitous across various paradigms for statistical inference. In particular, the conditional distribution or density of X given θ is usually referred to as 'the model.' What distinguishes the Bayesian approach, however, is that probability distributions are also used to describe available knowledge about the parameters, both before and after the data are observed. In particular, say the investigator selects a density function $f(\theta)$ to reflect his belief about the value of θ prior to observing the data. This so-called *prior distribution* can be interpreted in a rough sense as follows. If $f(\theta_1) > f(\theta_2)$ for parameter values θ_1 and θ_2, then the investigator views θ values close to θ_1 as being more plausible than θ values close to θ_2, with the ratio $f(\theta_2)/f(\theta_1)$ describing the strength of conviction in this assessment. More formally, prior distributions can be tied to axioms which link rational behaviour and preferences to the existence of *subjective probabilities*. Bernardo and Smith (1994) provide a comprehensive account of this, and other, axiomatic underpinnings of Bayesian analysis.

Armed with the statistical model and the prior distribution, one need only apply the laws of probability to determine the *posterior distribution* of the unknown θ given the observed data $X = x$. This process is summarized as *Bayes theorem*. Further using the common but lazy notation whereby 'f' denotes different densities depending on the context, Bayes theorem can be expressed as

$$f(\theta|x) \quad = \quad \frac{f(x|\theta)f(\theta)}{\int f(x|\theta^*)f(\theta^*)d\theta^*}, \tag{A.1}$$

where the denominator is simply the marginal density of X expressed in terms of the inputted model and prior densities. Under the Bayesian paradigm, (A.1) is a complete encapsulation of knowledge about the unknown θ after seeing the data x. Thus any inferential claims would be based entirely on the posterior distribution.

A useful insight about (A.1) is that the denominator of the right-hand side does not depend on θ. That is, it contributes nothing to the *shape* of the posterior distribution on θ; its only role is to ensure that this distribution is

properly normalized. Thus a simpler way to express Bayes theorem is

$$f(\theta|x) \quad \propto \quad f(x|\theta)f(\theta), \tag{A.2}$$

where it must be stressed that the proportionality is as a function of θ. In many situations it is simpler to work with (A.2) and drop multiplicative terms from $f(x|\theta)$ and $f(\theta)$ that do not depend on θ. In some simple problems one can normalize the posterior density by hand as a final step. In problems of more realistic complexity, the required integration cannot be done in closed-form. This necessitates the use of MCMC techniques or some other numerical scheme.

As a simple example of Bayes theorem in action, say an investigator wishes to estimate the prevalence θ of some exposure in a population. Prior to collecting any data, the investigator has some idea about the prevalence from sources such as anecdotal information, studies in similar populations, and scientific common sense. In the present scenario it is convenient to select the prior distribution from the Beta family of distributions. For completeness, the $Beta(c_1, c_2)$ distribution on the unit interval has density function

$$f(\theta|c_1, c_2) \quad = \quad \frac{\Gamma(c_1 + c_2)}{\Gamma(c_1)\Gamma(c_2)}\theta^{c_1-1}(1-\theta)^{c_2-1}. \tag{A.3}$$

In general when a prior is to be chosen from a parameterized family of distributions, the parameters of this family are termed *hyperparameters*, to distinguish them from the parameters θ in the statistical model for data. Now say the investigator's best prior guess at the prevalence is 0.25, but he acknowledges considerable uncertainty associated with this guess. Upon reflection, he assigns the prior distribution $\theta \sim Beta(2, 4)$. The mode of this distribution coincides with his best guess, as in general the mode of the $Beta(c_1, c_2)$ distribution is $(c_1 - 1)/(c_1 + c_2 - 2)$. But low values of (c_1, c_2) correspond to widely spread distribution, and indeed the plot of the $Beta(2, 4)$ density in Figure A.1 reflects the investigator's considerable *a priori* uncertainty about θ.

Using the standard notion of a sufficient statistic, the observable data can be summarized by the number of the n sampled subjects who are exposed, which we denote by X. Assuming independent sampling from a large population, we have $X|\theta \sim Binomial(n, \theta)$, i.e.

$$f(x|\theta) \quad = \quad \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x(1-\theta)^{n-x}. \tag{A.4}$$

Applying (A.3) and (A.4) in (A.2) gives the posterior density as

$$f(\theta|x) \propto \theta^{x+c_1-1}(1-\theta)^{n-x+c_2-1}.$$

Here we can identify that $f(\theta|x)$ is proportional to the $Beta(c_1 + x, c_2 + n - x)$ density, and therefore by normalization $f(\theta|x)$ must in fact be this Beta density function. Thus we have identified the denominator in (A.1) in an implicit manner. In the present context a Beta prior distribution is referred to as a *conjugate prior*, since a Beta prior distribution leads to a Beta posterior distribution.

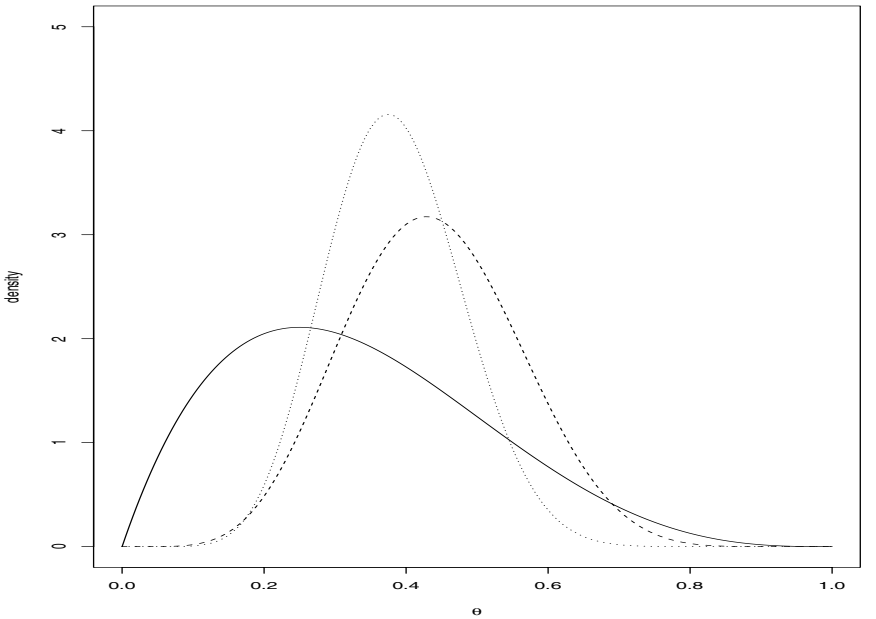

Figure A.1 *Prior and posterior densities of exposure prevalence. The solid curve is the Beta(2,4) prior density. The dashed curve is the Beta(7,9) posterior density after observing* 5 *out of* 10 *exposures in a sample from the population. The dotted curve is the Beta(10, 16) posterior density after observing* 8 *out of* 20 *exposures in a sample from the population.*

As a concrete example, say that in a sample of $n = 10$ subjects, $x = 5$ are exposed. The combination of these data with the investigator's prior leads to $\theta|x \sim Beta(7, 9)$, the density of which is given in Figure A.1. If three of a further ten sampled subjects are exposed, to yield $x = 8$ and $n = 20$, then $\theta|x \sim Beta(10, 16)$. This posterior density is also plotted in Figure A.1. After both $n = 10$ and $n = 20$ datapoints we see that the posterior mode 'compromises' between the prior mode of 0.25 and the sample proportion x/n, where the latter is the obvious 'data-only' estimate of θ. Note that when $n = 20$ the posterior distribution is narrower, and the prior distribution receives less 'weight' in the compromise than in the $n = 10$ scenario. These are very general phenomena: as the sample size increases the posterior narrows to reflect the accumulating evidence in the data, and the effect of the prior distribution on the posterior distribution diminishes. This is often expressed crudely as the 'data swamp the prior' as the sample size grows.

A.2 Point and Interval Estimates

While an overall summary of the post-data knowledge about an unknown parameter can be conveyed in the form of a posterior density, often specific point and interval estimates are desired. If a point estimate $\hat{\theta}$ of θ is needed, then intuitively one might take $\hat{\theta}$ to be a measure of where the posterior distribution of θ is centred. If the problem at hand yields a symmetric posterior distribution for the parameter in question, then the notion of centre is unambiguous. But say the posterior distribution in question were $\theta|\text{data} \sim Beta(2, 8)$. Then one might consider the posterior mean $\hat{\theta} = 2/10 = 0.2$, the posterior mode $\hat{\theta} = 1/8 = 0.125$, or the posterior median $\hat{\theta} \approx 0.180$ as a formal point estimate of θ. Informally, one might regard the choice between such estimators as a matter of taste. However, basic tenets of *decision theory* have something more precise to say on this matter.

Let $\hat{\theta}(x)$ be an arbitrary estimator of θ based on data x, and let $L(\theta, \hat{\theta})$ be a *loss function* which measures the loss incurred when θ is estimated by $\hat{\theta}$. A fundamental quantity of decision theory is the Bayes risk associated with an estimator and a prior distribution, defined as

$$r = \int\int L(\hat{\theta}(x), \theta) f(x|\theta) f(\theta) dx \, d\theta. \tag{A.5}$$

One way to think of the Bayes risk is as the average loss incurred when estimating θ in repeated experiments, where each experiment consists of 'nature' generating the true parameter value θ from the prior density $f(\theta)$ and then generating the data x from the model density $f(x|\theta)$. This differs from the standard frequentist 'thought experiment' whereby repeated datasets are generated under the same value of θ each time. By changing the order of integration in (A.5) it is easy to see that the estimator $\hat{\theta}(x)$ which minimizes the Bayes risk is

$$\hat{\theta}(x) = \text{argmin}_a \int L(\theta, a) f(\theta|x) d\theta, \tag{A.6}$$

where the integral on the right-hand side is the so called *posterior risk* associated with estimate or 'action' a. The fact that the optimal estimator is necessarily based on the posterior distribution via (A.6) is regarded as a sense in which Bayesian analysis is the optimal procedure in light of a prior weighting of the parameter space. Moreover, if one selects a loss function then there is no longer ambiguity about how to obtain a point estimate from the posterior distribution. For instance, it is easy to verify that if squared-error loss is chosen, so that $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, then the particular estimator which minimizes (A.6) is the posterior mean, $\hat{\theta}(x) = E(\theta|x)$. Similarly, under absolute error loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, (A.6) is minimized by the posterior median $\hat{\theta} = med(\theta|x)$.

Often interval estimates are also desired. This is particularly easy under the Bayesian paradigm, as the posterior distribution immediately conveys a sense of how likely it is that the parameter in question lies in a specific interval.

As a formal definition, any interval which is assigned probability $(1-\alpha)$ by the posterior distribution of θ is called a $100(1-\alpha)\%$ credible interval for θ. Note that the interpretation of a Bayesian credible interval is more direct than that of a frequentist confidence interval. Say, for instance, that (a, b) is a 95% credible interval for θ. Then it is legitimate to state the probability of θ lying between a and b given the observed data is 0.95. On the other hand, it is standard to caution that frequentist confidence intervals cannot be given this direct interpretation. The best that can be said is that on average such intervals contain the true parameter nineteen times out of twenty.

Of course for a given posterior distribution one can construct many different $100(1-\alpha)\%$ credible intervals. The two most common and intuitive choices are the equal-tailed credible interval and the highest posterior density (HPD) credible interval. As its name suggests, the $100(1-\alpha)\%$ equal-tailed credible interval has the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution $\theta|X$ as its endpoints. On the other hand, the $100(1-\alpha)\%$ HPD credible interval consists of all values of θ for which $f(\theta|x) > c$, where c is chosen to ensure that the interval does have posterior probability $1-\alpha$. In some situations involving a multimodal posterior distribution this definition could lead to an HPD region (say a union of disjoint intervals) rather than an interval. In the common situation of a unimodal posterior distribution, however, an interval is guaranteed. An important and easily verified property is that the $100(1-\alpha)\%$ HPD credible interval is the shortest possible $100(1-\alpha)\%$ credible interval.

It is often overlooked, but credible intervals do have a repeated sampling interpretation. As introduced with the notion of Bayes risk, consider repeated sampling whereby each sample is obtained by first generating θ from the prior distribution and then generating data x from the model distribution for $X|\theta$. It is trivial to establish that with respect to such sampling there is probability $(1-\alpha)$ that a $100(1-\alpha)\%$ credible interval for θ based on the X will actually contain θ. In this sense a credible interval is a weaker notion than a confidence interval which has the repeated sampling interpretation regardless of how the true θ values are generated. Of course in many practical problems it is not possible to construct exact confidence intervals, whereas the construction of credible intervals is entirely general.

A.3 Markov Chain Monte Carlo

In many problems the combination of the model distribution for data given parameters and the prior distribution for parameters leads to a complicated posterior distribution for parameters given data. By complicated we mean a nonstandard distribution under which it is hard to compute probabilities and expectations. More definitely, often there is no closed-form solution for the integral in the denominator of (A.1). The problem becomes particularly acute as the number of unknown parameters grows, as numerical integration techniques suffer from a *curse of dimensionality*. Until about 1990, these numerical difficulties severely limited the range of applications for Bayesian analysis.

The Bayesian approach to statistics underwent a quantum leap with the

realization that computational techniques originating in statistical physics and chemistry had great potential for determining posterior distributions arising in complex models with many parameters. The first algorithm to be applied in statistical contexts was the *Gibbs sampler* (Gelfand and Smith 1990, Gelfand *et al.* 1990), though this was soon realized to be just one specialization of the *Metropolis-Hastings* (MH) algorithm (Metropolis *et al.* 1953; Hastings 1970), and that other specializations were also very promising. Collectively the algorithms in question tend to be referred to as *Markov Chain Monte Carlo* (MCMC) algorithms. There are now numerous general references on MCMC methods, including the books of Gammerman (1997) and Chen, Shao, and Ibrahim (2000).

The broad idea is that one way to compute quantities associated with a particular *target distribution* is to simulate a large Monte Carlo sample of realizations from this distribution. Then probabilities and expectations under the target distribution can be approximated by analogous quantities in the Monte Carlo sample. Generally, if the Monte Carlo sample is comprised of m realizations, then the approximation error gets smaller at a rate of $m^{-1/2}$. When the target distribution is the posterior distribution over θ arising from a particular dataset, one can evaluate the unnormalized version of the target density (A.2) at any given θ value. The challenge is to use such evaluations to draw a Monte Carlo sample from the target distribution.

Usually with a complicated likelihood function and prior density one cannot (a) simulate independent and identically distributed realizations from the posterior distribution, or even (b) simulate *dependent* but identically distributed realizations from the posterior distribution. What MCMC methods can do, however, is (c) simulate from a *Markov chain* having the posterior distribution as its stationary distribution. Theory dictates that under weak regularity conditions a Markov chain will converge quickly to its stationary distribution, so in fact (c) is almost as good as (b). One can simulate from the Markov chain and then discard the first portion of sampled values as a 'burn-in' period which allows the chain to converge. The remaining sampled values are then treated as (b), a dependent sample from the posterior distribution. In general one can still estimate features of a target distribution with a dependent rather than independent sample from the distribution. However, this estimation becomes more inefficient as the strength of the serial dependence in the Markov chain increases. Hence there are two different concerns about how well a particular MCMC algorithm works in learning the features of a particular posterior distribution: is convergence to the posterior distribution fast, and does the sampler *mix* or move around the possible parameter values rapidly. That is, is the dependence mild enough for the Monte Carlo sample to yield good estimates of posterior distribution features.

The MH algorithm has a remarkably simple form. Say the target density of interest for a vector θ of dimension p is $s(\theta)$, and let $t(b|a)$ be a conditional distribution from which we can simulate a *candidate state* b given a *current state* a. We generate our Markov chain $\theta^{(1)}, \ldots, \theta^{(m)}$ by setting $\theta^{(1)}$ to have

some arbitrary value, and then iterating the following three-part updating scheme to generate $\theta^{(i+1)}$ given $\theta^{(i)}$.

1. Simulate a candidate value θ^* from the conditional density $t(\theta^*|\theta^{(i)})$.
2. Compute the acceptance probability

$$pr = \min\left\{\frac{s(\theta^*)}{s(\theta^{(i)})}\frac{t(\theta^{(i)}|\theta^*)}{t(\theta^*|\theta^{(i)})}, 1\right\}.$$

3. Simulate a 'biased coin flip' to set

$$\theta^{(i+1)} = \begin{cases} \theta^* & \text{with probability } pr, \\ \theta^{(i)} & \text{with probability } 1 - pr. \end{cases}$$

Note that only an unnormalized expression for the target density is required to implement the MH algorithm, as the acceptance probability depends on s only through the ratio of $s(\theta^*)/s(\theta^{(i)})$. Clearly when the target density is a posterior density this ratio is $\{f(y|\theta^*)f(\theta^*)\}/\{f(y|\theta^{(i)})f(\theta^{(i)})\}$, which can be evaluated regardless of whether the posterior density can be normalized in closed-form.

It it quite straightforward to show that the Metropolis-Hastings update does have the target density as its stationary distribution. That is, if $\theta^{(i)} \sim s()$ then $\theta^{(i+1)} \sim s()$. Markov chain theory then dictates that under weak regularity conditions the distribution of $\theta^{(i)}$ converges quickly to $s()$ as i increases. Much of the requisite Markov chain theory is made quite accessible by Robert and Casella (1999). We emphasize that here 'quick' convergence means at an exponential rate, so that formal measures of the distance between the actual distribution of $\theta^{(i)}$ and the target distribution behave like $\exp(-ci)$ for some $c > 0$, as opposed to the slow $i^{-1/2}$ convergence of most limiting theorems in statistics. Thus after allowing m^* burn-in iterations one can estimate target distribution features with corresponding sample features of $\theta^{(m^*+1)}, \ldots, \theta^{(m^*+m)}$. An important aspect of MCMC is choosing m^* and m to be large enough to compute target distribution features with desired accuracy.

The MH algorithm is exceedingly general in that one can choose any conditional density $t(\cdot|\cdot)$ from which to generate candidates. In practice, however, one must make intelligent choices in order to get reasonable performance. Often it is necessary to reduce a high-dimensional problem to a series of one-dimensional problems, by generating candidate vectors which differ from the current vector in only one component.

The Gibbs sampler is such an algorithm that is useful when the joint target distribution on a vector θ does not have a standard form but does yield standard forms for the distribution of each single component given the other components. The target distribution of a single component given all the other components is referred to as the *full conditional* distribution of that component. It is easy to check that generating a candidate by sampling from the full conditional for one component while leaving the other components unchanged leads to an acceptance probability of one in the MH algorithm. Thus candidates are never rejected. To be more specific, if all three full conditionals for

$\theta = (\theta_1, \theta_2, \theta_3)$ are standard distributions, then one complete Gibbs sampler iteration from $\theta^{(i)}$ to $\theta^{(i+1)}$ is completed as

1. simulate $\theta_1^{(i+1)}$ from $s(\theta_1 | \theta_2 = \theta_2^{(i)}, \theta_3 = \theta_3^{(i)})$,
2. simulate $\theta_2^{(i+1)}$ from $s(\theta_2 | \theta_1 = \theta_1^{(i+1)}, \theta_3 = \theta_3^{(i)})$,
3. simulate $\theta_3^{(i+1)}$ from $s(\theta_3 | \theta_1 = \theta_1^{(i+1)}, \theta_2 = \theta_2^{(i+1)})$.

The obvious drawback to the Gibbs sampler is that some full conditionals may be very difficult to simulate from. While there are some fairly general approaches to so doing (see, for instance, Gilks and Wild 1992; Gilks, Best and Tan 1995), sometimes candidate generation schemes which do not rely on sampling the full conditional distributions are necessary. One of the simplest such schemes is the *random-walk* MH algorithm. If the strategy of reducing to one-dimensional updates is followed, a candidate θ^* is obtained from $\theta^{(i)}$ by adding 'noise' to one component of θ^*. By generating the noise from a distribution which is symmetric about zero, we obtain the simplification that $t(\theta^* | \theta^{(i)}) = t(\theta^{(i)} | \theta^*)$. Consequently the acceptance probability will depend only on the ratio $s(\theta^*)/s(\theta^{(i)})$, making it very simple to implement this algorithm. Typically the primary challenge is 'tuning' the variation in the noise distribution. In particular, say that when the j-th component of θ is updated the candidate is formed via $\theta_j^* \sim N\left(\theta_j^{(i)}, \sigma_j^2\right)$. One must then select values of $\sigma_1, \ldots, \sigma_p$ which yield reasonable performance from the algorithm, where σ_j is often referred to as the *jump size* for the θ_j update. To large extent this is a trial-and-error process, though there is some useful guidance in the literature to which we refer shortly.

Perhaps the vast majority of MCMC-based inference done in practice uses some combination of the Gibbs sampler and the random walk MH algorithms, primarily because they are simple to understand and implement. On the other hand, more specialized and powerful MH algorithms are sometimes needed. Numerous such algorithms have been investigated, such as the hybrid algorithm which uses extra information in the form of derivatives of the posterior density, and various algorithms which introduce auxiliary variables in an attempt to more freely traverse the parameter space. Also, parameter reparameterizations which tend to improve MCMC performance have been studied. Recent books which cover more specialized algorithms include Chen, Shao, and Ibrahim (2000), and Liu (2001).

To get some intuition for MCMC we try the Gibbs sampler and random walk MH algorithms on a simple bivariate normal target distribution, namely

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Of course this is an artificial example, as it is trivial to sample directly from the target distribution. Nonetheless, we will be able to develop some intuition about the algorithms in this test case.

Figure A.2 gives traceplots of θ_1 for 500 iterations of the Gibbs sampler and of the random-walk MH algorithm with three different jump sizes

($\sigma = 0.25$, $\sigma = 1.25$, $\sigma = 3$). In fact two different target distributions are considered ($\rho = 0.5$, $\rho = 0.9$). In each case the initial state is taken to be $(\theta_1^{(1)}, \theta_2^{(1)}) = (10, 10)$, which of course is very far from where the target distribution is concentrated. All four algorithms appear to mix better for the $\rho = 0.5$ target distribution than for the $\rho = 0.9$ target distribution. In fact this is a very general phenomenon: MCMC algorithms tend to perform less well when the target distribution involves stronger dependencies. For both target distributions the Gibbs sampler outperforms the random-walk MH algorithm, in terms of both faster convergence and better mixing. This is not surprising given that the Gibbs sampler relies on more knowledge about the target distribution in the form of the full conditional distributions. Thus Gibbs sampling tends to be favoured whenever full conditional distributions have standard forms.

It is clear from Figure A.2 that the performance of the random walk MH algorithm depends quite strongly on the choice of jump size. Intuitively, the jump size will correlate negatively with the *acceptance rate*, i.e., the proportion of the iterations at which the candidate state is accepted. A small jump size will correspond to small differences between the target density at the current and candidate states, so that small values of the acceptance probability are unlikely. Conversely, a large jump size will tend to generate candidate states which miss the region where the target distribution is concentrated. The form of the acceptance probability indicates that such candidates are very likely to be rejected. The acceptance rates given in Figure A.2 do indeed decrease as the jump size increases.

On the face of it, a high acceptance rate seems good, as it suggests the sampler is unlikely to get stuck at the same value for a number of successive iterations. However, this comes at the cost of only making very small moves. Thus with a jump size of $\sigma = 0.25$ (yielding acceptance rates of 85% when $\rho = 0.5$ and 78% when $\rho = 0.9$) the algorithm is seen to converge slowly and mix poorly in Figure A.2. The figure indicates that the tradeoff of bigger jumps ($\sigma = 1.25$) at the cost of lower acceptance rates (51% when $\rho = 0.5$ and 33% when $\rho = 0.9$) seems worthwhile. That is, both faster convergence and better mixing are evident. Intuitively, however, a very low acceptance rate does not seem appealing, as this would entail the value of θ not changing at most iterations. This is supported by the poor 'boxy' mixing with a jump size of $\sigma = 3$ (and acceptance rates of 24% when $\rho = 0.5$ and 13% when $\rho = 0.9$). Indeed, theoretical work of Roberts, Gelman, and Gilks (1997) supports the notion that the random-walk MH algorithm works best at a mid-range acceptance rate, with common practice being to select jump sizes by trial-and-error in order to obtain an acceptance rate perhaps in the range of 35% to 60%. Roberts and Rosenthal (2001) provide an accessible review of work on the so called 'optimal scaling' of MH algorithms.

As a final comment about the example, note that a particular jump size yields a lower acceptance rate when ρ is larger. Since we are applying the random walk MH algorithm in a 'one-component at a time' fashion, the relevant

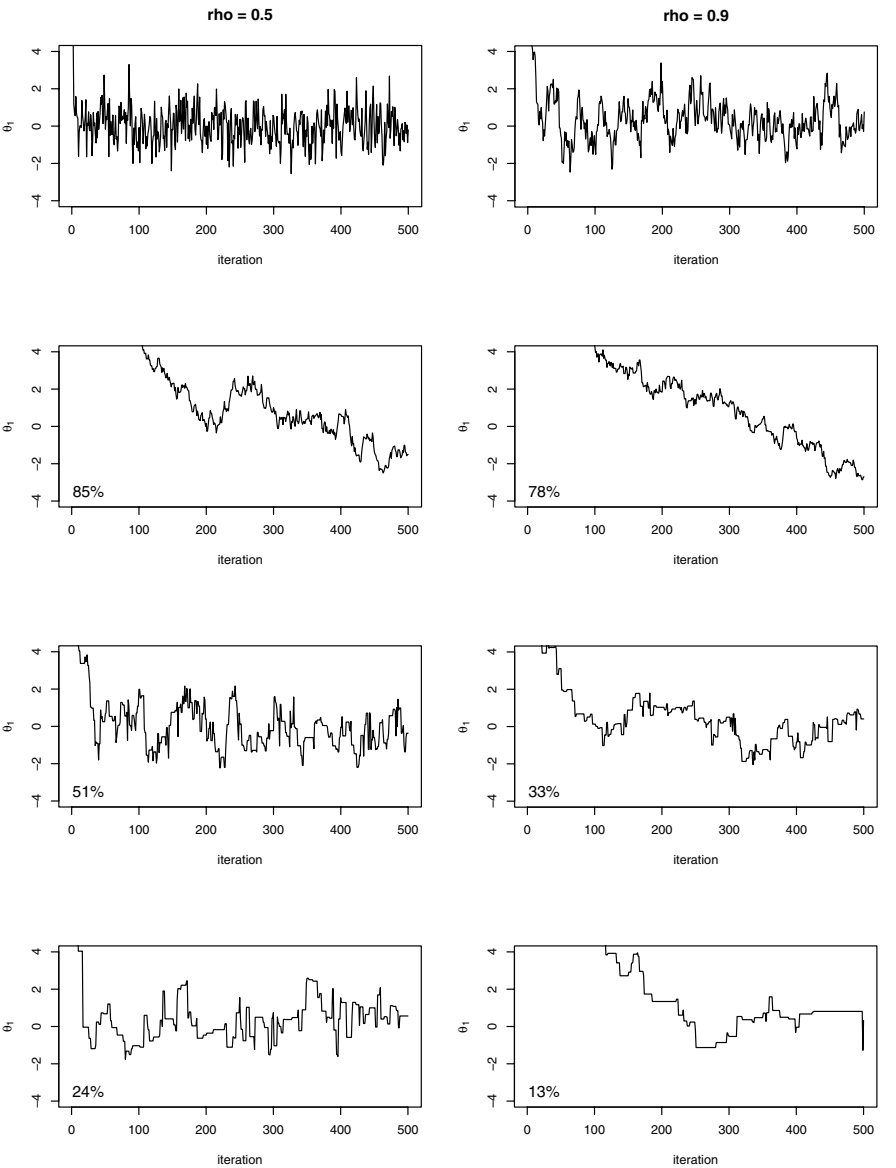

Figure A.2 *MCMC algorithms applied to a bivariate normal target distribution. Each panel plots the sampled values of* θ_1 *against iteration, for 500 iterations of the algorithm. The left panels correspond to a* $\rho = 0.5$ *target, while the right panels correspond to* $\rho = 0.9$. *From top to bottom the rows correspond to different algorithms: the Gibbs sampler, and the random walk MH algorithm with jump sizes* $\sigma = 0.25$, $\sigma = 1.25$, $\sigma = 3$. *For the random walk MH algorithm the acceptance rate is given in the bottom left corner of each panel.*

variation in the posterior distribution is the variation in the full conditional distributions. That is, the update to the j-th component of θ can be regarded as an update with the j-th full conditional distribution as the target distribution. Since larger ρ corresponds to narrower full conditionals in our example, a fixed jump size σ can be thought of as becoming effectively larger as ρ increases.

Here we have examined the MCMC output very informally, making qualitative statements about how good the output is, and consequently how well it might represent the target distribution. In fact there is a large literature on more formal assessments. Much of this revolves around *convergence diagnostics*, which attempt to determine empirically how many burn-in iterations are required. The first and perhaps most popular diagnostic is that of Gelman and Rubin (1992). It involves a simple scheme based on implementing multiple chains and comparing between and within chain variation. Much of the later suggestions are more technical, involving more sophisticated Markov chain theory. Many of the suggested schemes are reviewed by Mengersen, Robert and Guihenneuc-Jouyaux (1999).

As a further point about MCMC output for a given posterior distribution, usually the actual computation of point and interval estimates is carried out in the intuitively obvious way. For instance, say that post burn-in output $\theta^{(1)}, \ldots, \theta^{(m)}$ is obtained when the target distribution is a posterior distribution for θ given observed data. And say the scalar parameter $\psi = g(\theta)$ is of interest. Then the sample mean or median of $g(\theta^{(1)}), \ldots, g(\theta^{(m)})$ could be reported as the posterior mean or median of ψ. Similarly, appropriate sample quantiles could be reported as an equal-tailed credible interval for ψ, while the computation of HPD credible intervals is slightly more involved. We emphasize that these are actually *approximations* to posterior quantities which involve numerical error, often referred to as Monte Carlo error or simulation error. The exact values of posterior quantities would only be obtained in the large m limit of infinite Monte Carlo sample size. However, typically the approximation error can be made very small in relation to the actual statistical uncertainty at play. With a careful choice of MCMC algorithm and a sufficiently large m, for instance, the difference between the Monte Carlo approximation and the actual posterior mean of a parameter can be made as small as desired in relation to the difference between the actual posterior mean and the true value of the parameter. Following along these lines, sometimes it is relevant to report a simulation standard error, perhaps on the basis of the variation in estimates over multiple Markov chain runs. Then the relevant check would be whether the simulation standard error is indeed small in relation to say the posterior standard deviation of the parameter in question.

Finally, often one wishes to convey the shape of marginal posterior distributions, either instead of, or in addition to, formal point and interval estimates. With respect to the notation above, an obvious approximation to the posterior marginal distribution of ψ is a histogram of $g(\theta^{(1)}), \ldots, g(\theta^{(m)})$. If the inherent coarseness of a histogram is not appealing, then a smooth approxi-

mation to the marginal posterior density can be obtained by applying kernel density estimation to the sampled values. As a default in this book we have used kernel density estimation with the general-purpose bandwidth suggested by Silverman (1986). On occasion, however, it is necessary to increase the bandwidth by hand, in order to obtain a plausibly smooth approximation to the posterior marginal density.

A.4 Prior Selection

A necessity with any Bayesian analysis is the specification of a prior distribution for the unknown parameters. Depending on one's views and the nature of the problem at hand, this can be viewed as a strength or a weakness of the Bayesian approach. A 'pure' Bayesian approach dictates that the prior distribution is strictly a representation of the investigator's belief about the parameters before seeing the data, and a sizeable literature describes *elicitation procedures* to capture beliefs in the form of probability distributions. Some recent references on elicitation include Kadane and Wolfson (1998), and O'Hagan (1998). It is the case, however, that most Bayesian analysis done in practice makes limited or no use of formal elicitation procedures. This is probably because meaningful elicitation is viewed as too difficult, particularly when the number of unknown parameters is large. Instead, often computational convenience and a desire to *not* impart strong prior views into the analysis are primary factors in the selection of a prior distribution.

As a simple example, say $X_1, \ldots, X_n$ are independent and identically distributed from a $N(\mu, \sigma^2)$ distribution. For computational expedience one might consider a prior distribution of the form

$$\begin{aligned} \mu|\sigma^2 &\sim N\left(\tilde{\mu}, \frac{\sigma^2}{k_1}\right), \\ \sigma^2 &\sim IG\left(\frac{k_2}{2}, \frac{k_2\tilde{\sigma}^2}{2}\right), \end{aligned}$$

where N and IG denote *normal* and *inverse-gamma* distributions respectively. Note that $(\tilde{\mu}, \tilde{\sigma}^2, k_1, k_2)$ are hyperparameters that must be specified. In particular, $\tilde{\mu}$ and $\tilde{\sigma}^2$ can be regarded as prior guesses at the parameter values, particularly as an $IG(a, b)$ distribution is roughly centred at b/a. The role of k_1 and k_2 is less clear, but we will see momentarily that intuitive meanings can be ascribed to these hyperparameters. Another curious feature of this prior is that the two parameters are *not* independent *a priori*. The rationale for this derives more from the resultant interpretability of the posterior distribution than from an inherent sense that prior beliefs about the two parameters ought to be correlated. An important feature is that the prior is *conditionally conjugate*, in the sense that $\mu|\sigma^2, x$ has a normal distribution while $\sigma^2|\mu, x$ has an inverse gamma distribution. Consequently, it is trivial to implement Gibbs sampling in such a scenario, and in more complicated variants as well.

More specifically, the posterior full conditional distributions are identified

as

$$\mu|\sigma^2, x \quad \sim \quad N\left\{\left(\frac{n}{n+k_1}\right)\bar{x} + \left(\frac{k_1}{k_1+n}\right)\tilde{\mu}, \frac{\sigma^2}{n+k_1}\right\} \tag{A.7}$$

and

$$\sigma^2|\mu, x \quad \sim \quad IG\left(\frac{n+k_2}{2}, \frac{ns^2+k_2\tilde{\sigma}^2}{2}\right), \tag{A.8}$$

where $\bar{x}$ and s^2 denote the sample mean and variance. These forms suggest k_1 and k_2 be interpreted as *effective sample sizes* for the prior distributions of μ and σ^2 respectively. This is evident from examining both the centre and variation of the distributions (A.7) and (A.8). We see that both distributions are centred at linear combinations of a sample-based estimate and the prior guess, with the former having weight $n/(n+k_j)$ and the latter having weight $k_j/(k_j+n)$. Moreover, the variation in both distributions declines as $1/(n+k_j)$. Thus the information content in prior (A.7) can be viewed as equivalent to k_1 observed data points, and the content of (A.8) as equivalent to k_2 observed data points. A relatively common approach is to choose $k_1 = 1$ and $k_2 = 1$, to ensure the prior gets little weight compared to the data. Such a choice has been termed a *unit-information prior* by Kass and Wasserman (1995), who investigate some theoretical aspects of using such priors. Thinking about prior specification in terms of prior guesses and effective sample sizes is a pragmatic way of obtaining a reasonable prior distribution without an extensive elicitation process.

No matter what degree of comfort one has in a particular prior, it seems reasonable to ask how posterior inferences would change if the prior were changed. Indeed, some would argue that such *prior sensitivity* analysis is a vital part of the Bayesian approach. Thus it is relatively common to see an analysis repeated with a few different choices of prior distribution. As well, more formal mathematical techniques can be used to quantify the sensitivity of a Bayesian inference to the choice of prior. Ríos-Insua and Ruggeri (2000) provide a recent overview of such techniques.

A.5 MCMC and Unobserved Structure

Much of the advantage of MCMC-based inference arises in problems where succinct expression of the model distribution for data given parameters, or the prior distribution for parameters, is problematic. In many problems there is inherent unobserved structure that in some sense lies between the observed data and the unknown parameters. Outside the realm of mismeasured variables, examples of such unobserved quantities include the random effects in a variance components model, the actual failure times for censored observations in a lifetime data model, the unobserved continuous responses linked to observed binary responses in a probit model, and the values of missing covariates in a regression model.

To be more specific, consider a model involving random effects. Let A de-

note the data to be observed, let B be the unobservable random effects, and let C be the fixed effects and variance components. Traditionally one would regard only C as unknown parameters, so that Bayesian inference would be based on the distribution of $C|A$. On the face of it one would require a model distribution for $(A|C)$ and a prior distribution for C in order to obtain this posterior distribution. However, depending on the form of the random effects model it may not be possible to determine the distribution of $(A|C)$ explicitly. An appealing alternative is to consider a joint distribution over all quantities, in the form of $f(a,b,c) = f(a|b,c)f(b|c)f(c)$. Then Bayes theorem and MCMC allow one to obtain a posterior sample from $(B,C|A)$, the joint distribution of random effects and unknown parameters given observed data. Then trivially the sampled C values comprise a sample from the distribution of $(C|A)$, enabling the desired inferences. Simply put, MCMC is very effective at exploiting complex hierarchical structure in models, with the requisite integration carried out behind the scenes.

Another example that falls into the framework above is models where the response variable can be represented via latent structure. For instance, the binary responses A in a probit model can be viewed as arising from thresholding of unobserved continuous responses B, with the relevant parameters again denoted by C. While there is no closed-form for the posterior distribution of parameters C given data A, there is a closed-form for the posterior distribution of parameters C given A and B. Thus this full conditional distribution is readily sampled, as are the full conditionals for A and B. In essence MCMC makes this intractable problem tractable, as noted by Carlin and Polson (1992), and Albert and Chib (1993).

In a slightly more elaborate example of direct relevance to this book, let A be the mismeasured values of an explanatory variable, let B be the response variable values, let C be the true but unobservable values of the explanatory variable, and let D be the unknown parameters involved. Ultimately one wishes to base inferences on the distribution of $(D|A,B)$; however it may be impossible to get an explicit form for $(A,B|D)$ in order to apply Bayes theorem directly. Instead, one can use MCMC to sample from $(C,D|A,B)$, given that the distributions of $(A|B,C,D)$, $(B|C,D)$, $(C|D)$, and D are available. As above, an ability to sample from $(C,D|A,B)$ automatically implies an ability to sample from the desired distribution of $(D|A,B)$. Again MCMC can be very effective at exploiting the model structure to get at the desired but complicated distributions via simpler distributions.