Contents

Chapter 6

Approximations

6.1 The grid method

When one is confronted with a low-dimensional problem with a continuous parameter, then it is usually easy to approximate the posterior density on a dense grid of points which covers the relevant part of the parameter space. We discuss the method for a one-dimensional parameter θ.

We suppose that the posterior is available in the unnormalized form

$$f_{\Theta|Y}(\theta \mid y) = \frac{1}{c(y)}\, q(\theta \mid y),$$

where we know how to evaluate the unnormalized density $q(\theta \mid y)$, but do not necessarily know the value of the normalizing constant $c(y)$.

Instead of the original parameter space, we consider a finite interval $[a, b]$, which should cover most of the mass of the posterior distribution. We divide $[a, b]$ evenly into N subintervals

$$B_i = [a + (i-1)h, a + ih], \qquad i = 1, \ldots, N.$$

The width h of one subinterval is

$$h = \frac{b-a}{N}.$$

Let θ_i be the midpoint of the i'th subinterval,

$$\theta_i = a + (i - \frac{1}{2})h, \qquad i = 1, \ldots, N.$$

We use the midpoint rule for numerical integration. This means that we approximate the integral over the i'th subinterval of any function g by the rule

$$\int_{B_i} g(\theta)\, \mathrm{d}\theta \approx h g(\theta_i). \tag{6.1}$$

Using the midpoint rule on each of the subintervals, we get the following

approximation for the normalizing constant

$$\begin{aligned} c(y) = \int q(\theta \mid y)\,\mathrm{d}\theta &\approx \int_a^b q(\theta \mid y)\,\mathrm{d}\theta = \sum_{i=1}^{N} \int_{B_i} q(\theta \mid y)\,\mathrm{d}\theta \\ &\approx h \sum_{i=1}^{N} q(\theta_i \mid y) \end{aligned} \tag{6.2}$$

Using this approximation, we can approximate the value of the posterior density at the point θ_i,

$$f_{\Theta|Y}(\theta_i \mid y) = \frac{1}{c(y)}\, q(\theta_i \mid y) \approx \frac{1}{h} \frac{q(\theta_i \mid y)}{\sum_{j=1}^{N} q(\theta_j \mid y)}. \tag{6.3}$$

We also obtain approximations for the posterior probabilities of the subintervals,

$$\begin{aligned} P(\Theta \in B_i \mid Y = y) = \int_{B_i} f_{\Theta|Y}(\theta \mid y)\,\mathrm{d}\theta &\approx h f_{\Theta|Y}(\theta_i \mid y) \\ &\approx \frac{q(\theta_i \mid y)}{\sum_{j=1}^{N} q(\theta_j \mid y)}. \end{aligned} \tag{6.4}$$

By following the same reasoning which lead to (6.2), we may form the approximation

$$\int k(\theta)\, q(\theta \mid y)\,\mathrm{d}\theta \approx h \sum_{i=1} k(\theta_i)\, q(\theta_i \mid y)$$

basically for any function k such that $k(\theta)\, q(\theta \mid y)$ differs appreciably from zero only on the interval (a, b). This can be used to approximate the posterior expection of an arbitrary function $k(\theta)$ of the parameter, by

$$\begin{aligned} E(k(\Theta) \mid Y = y) = \int k(\theta)\, f_{\Theta|Y}(\theta \mid y)\,\mathrm{d}\theta &= \frac{\int k(\theta)\, q(\theta \mid y)\,\mathrm{d}\theta}{\int q(\theta \mid y)\,\mathrm{d}\theta} \\ &\approx \frac{\sum_{i=1}^{N} k(\theta_i)\, q(\theta_i \mid y)}{\sum_{j=1}^{N} q(\theta_j \mid y)} \end{aligned} \tag{6.5}$$

These approximations can be surprisingly accurate even for moderate values of N provided we are able to identify an interval $[a, b]$, which covers the essential part of posterior distribution.

To summarize, the grid method for approximating the posterior density or for or simulating from it is the following.

- First evaluate the unnormalized posterior density $q(\theta \mid y)$ at a regular grid of points $\theta_1, \ldots, \theta_N$ with spacing h. The grid should cover the main support of the posterior density.
- If you want to plot the posterior density, normalize these values by dividing by their sum and additionally by the bin width h as in eq. (6.3). This gives an approximation to the posterior ordinates $p(\theta_i \mid y)$ at the grid points θ_i.
- If you want a sample from the posterior, sample with replacement from the grid points θ_i with probabilities proportional to the numbers $q(\theta_i \mid y)$, cf. (6.4).

- If you want to approximate the posterior expectation $E[k(\theta) \mid y]$, calculate the weighted average of the values $k(\theta_i)$ using the values $q(\theta_i \mid y)$ as weights, cf. eq. (6.5).

The midpoint rule is considered a rather crude method of numerical integration. In the numerical analysis literature, there are available much more sophisticated methods of numerical integration (or numerical quadrature) and they can be used in a similar manner. Besides dimension one, these kinds of approaches can be used in dimensions two or three. However, as the dimensionality of the parameter space grows, computing at every point in a dense multidimensional grid becomes more and more expensive.

6.2 Normal approximation to the posterior

We now try to approximate a posterior density by a normal density based on the behavior of the posterior density at its mode. This approximation can be quite accurate, when the sample sizes is large, provided the posterior is unimodal. We will call the resulting approximation a normal approximation to the posterior, but the result is sometimes also called a Laplace approximation or a modal approximation. A normal approximation can be used directly as an approximate description of the posterior. However, such an approximation can be utilized also indirectly, e.g., to form a good proposal distribution for the Metropolis–Hastings method.

We first discuss normal approximation in the univariate situation. The statistical model has a single parameter θ, which has a continuous distribution. We do know an unnormalized version $q(\theta \mid y)$ of the posterior density, but the normalizing constant is usually unknown. We consider the case, where $\theta \mapsto q(\theta \mid y)$ is unimodal: i.e., it has only one local maximum. We suppose that we have located the mode $\hat{\theta}$ of the unnormalized posterior $q(\theta \mid y)$. Notice that $\hat{\theta}$ is also the posterior mode, which is also called the MAP (maximum a posteriori) estimate. Actually, $\hat{\theta}$ depends on the data y, but we suppress this dependence in our notation. Usually we would have to run some numerical optimization algorithm in order to find the mode.

The basic idea of the method is to use the second degree Taylor polynomial of the log-posterior (the logarithm of the posterior density) centered on the mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + b(\theta - \hat{\theta}) - \frac{1}{2}A(\theta - \hat{\theta})^2, \tag{6.6}$$

where

$$b = \frac{\partial}{\partial\theta} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = 0,$$

and

$$A = -\frac{\partial^2}{\partial\theta^2} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial\theta^2} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}}.$$

Notice the following points.

- The first and higher order (partial) derivatives with respect to θ of $\log q(\theta \mid y)$ and $\log f_{\Theta|Y}(\theta \mid y)$ agree, since these function differ only by an additive constant (which depends on y but not on θ).

- The first order term of the Taylor expansion disappears, since $\hat{\theta}$ is also the mode of the log-posterior $\log f_{\Theta|Y}(\theta \mid y)$.
- $A \geq 0$, since $\hat{\theta}$ is a maximum of $q(\theta \mid y)$. For the following, we need to assume that $A > 0$.

Taking the exponential of the second degree Taylor approximation (6.6), we see that we may approximate the posterior by the function

$$\pi_{\text{approx}}(\theta) \propto \exp\left(-\frac{A}{2}(\theta - \hat{\theta})^2\right),$$

at least in the vicinity of the mode $\hat{\theta}$. Luckily, we recognize that $\pi_{\text{approx}}(\theta)$ is an unnormalized form of the density of the normal distribution with mean $\hat{\theta}$ and variance $1/A$. The end result is that the posterior distribution can be approximated with the normal distribution

$$N\left(\hat{\theta}, \frac{1}{-L''(\hat{\theta})}\right), \tag{6.7}$$

where $L(\theta)$ is the logarithm of the unnormalized posterior,

$$L(\theta) = \log q(\theta \mid y)$$

and $L''(\hat{\theta})$ is the second derivative of $L(\theta)$ evaluated at the mode $\hat{\theta}$.

The multivariate analog of the result starts with the second degree expansion of the log-posterior centered on its mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + 0 - \frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta}),$$

where A is the negative Hessian matrix of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode,

$$A_{ij} = -\left.\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_{\Theta|Y}(\theta \mid y)\right|_{\theta=\hat{\theta}} = -\left.\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\theta)\right|_{\theta=\hat{\theta}} = -\left[\left.\frac{\partial^2}{\partial\theta\, \partial\theta^T} L(\theta)\right|_{\theta=\hat{\theta}}\right]_{ij}$$

The first degree term of the expansion vanishes, since $\hat{\theta}$ is the mode of the log-posterior. Here A is at least positively semidefinite, since $\hat{\theta}$ is a maximum. If A is positively definite, we can proceed with the normal approximation.

Exponentiating, we find out that approximately (at least near the mode)

$$f_{\Theta|Y}(\theta \mid y) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta})\right).$$

Therefore we can approximate the posterior with the corresponding multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix given by A^{-1}, i.e., the approximating normal distribution is

$$N\left(\hat{\theta}, \left(-L''(\hat{\theta})\right)^{-1}\right), \tag{6.8}$$

where $L''(\hat{\theta})$ is the Hessian matrix of the logarithm of the unnormalized posterior, $L(\theta) = \log q(\theta \mid y)$, evaluated at its mode $\hat{\theta}$. The precision matrix of the

approximating normal distribution is the negative Hessian of the log-posterior evaluated at the posterior mode. Another characterization for the precision matrix is that it is the Hessian of the negative log-posterior evaluated at the posterior mode. The covariance matrix of the normal approximation is the inverse of its precision matrix.

Typically the mode of the log-posterior (or the maximum point of the negative log-posterior) would be calculated using some numerical optimization algorithm. The Hessian would then be calculated using numerical differentiation, see Sec. B.7 for an example.

Before using the normal approximation, it is often advisable to reparameterize the model so that the transformed parameters are defined on the whole real line and have roughly symmetric distributions. E.g., one can use logarithms of positive parameters and apply the logit function to parameters which take values on the interval $(0, 1)$. The normal approximation is then constructed for the transformed parameters, and the approximation can then be translated back to the original parameter space. One must, however, remember to multiply by the appropriate Jacobians.

Example 6.1. We consider the unnormalized posterior

$$q(\theta \mid y) = \theta^{y_4} \, (1-\theta)^{y_2+y_3} \, (2+\theta)^{y_1}, \qquad 0 < \theta < 1,$$

where $y = (y_1, y_2, y_3, y_4) = (13, 1, 2, 3)$. The mode and the second derivative of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode are given by

$$\hat{\theta} \approx 0.677, \qquad L''(\hat{\theta}) \approx -37.113.$$

(The mode $\hat{\theta}$ can be found by solving a quadratic equation.) The resulting normal approximation in the original parameter space is $N(0.677, 1/37.113)$.

We next reparametrize by defining ϕ as the logit of θ,

$$\phi = \operatorname{logit}(\theta) = \ln \frac{\theta}{1-\theta} \quad \Leftrightarrow \quad \theta = \frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}}.$$

The given unnormalized posterior for θ transforms to the following unnormalized posterior for ϕ,

$$\begin{aligned} \tilde{q}(\phi \mid y) &= q(\theta \mid y) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right| \\ &= \left(\frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_4} \left(\frac{1}{1+\mathrm{e}^{\phi}} \right)^{y_2+y_3} \left(\frac{2+3\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_1} \frac{\mathrm{e}^{\phi}}{(1+\mathrm{e}^{\phi})^2}. \end{aligned}$$

The mode and the second derivative of $\tilde{L}(\phi) = \log \tilde{q}(\phi \mid y)$ evaluated at the mode are given by

$$\hat{\phi} \approx 0.582, \qquad \tilde{L}''(\hat{\phi}) \approx -2.259.$$

(Also $\hat{\phi}$ can be found by solving a quadratic.) This results in the normal approximation $N(0.582, 1/2.259)$ for the logit of θ.

When we translate that approximation back to the original parameter space, we get the approximation

$$f_{\Theta|Y}(\theta \mid y) \approx N(\phi \mid 0.582, 1/2.259) \left| \frac{\mathrm{d}\phi}{\mathrm{d}\theta} \right|,$$

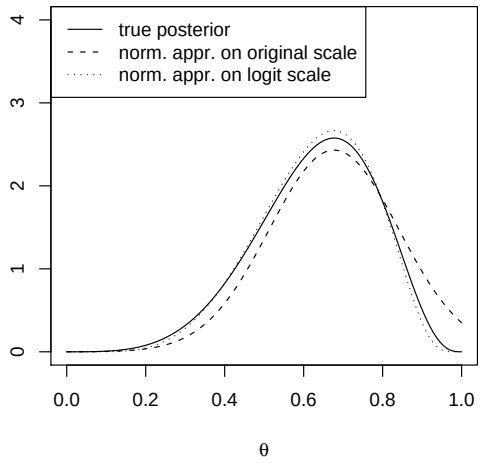


Figure 6.1: The exact posterior density (solid line) together with its normal approximation (dashed line) and the approximation based on the normal approximation for the logit of θ. The last approximation is markedly non-normal on the original scale, and it is able to capture the skewness of the true posterior density.

i.e.,

$$f_{\Theta|Y}(\theta \mid y) \approx N(\mathrm{logit}(\theta) \mid 0.582, 1/2.259)\, \frac{1}{\theta(1-\theta)}.$$

Both of these approximations are plotted in Figure 6.1 together with the true posterior density (whose normalizing constant can be found exactly). $\triangle$

We finish by discussing the relationship of the normal approximation (6.8) to the frequentist asymptotics of the maximum likelihood estimator. The unnormalized version of the posterior density is of the form

$$q(\theta \mid y) = k(y)\, f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta) = k(y)\, p(y \mid \theta)\, p(\theta),$$

where $p(\theta)$ is the prior, $p(y \mid \theta)$ is the likelihood, and $k(y)$ is any convenient constant which may depend on the data but not on the parameter vector. Therefore the logarithm of the unnormalized posterior is

$$L(\theta) = \log q(\theta \mid y) = \log k(y) + \ell(\theta) + \log p(\theta),$$

where $\ell(\theta) = p(y \mid \theta)$ is the log-likelihood. Therefore the negative Hessian of $L(\theta)$ is

$$-L''(\theta) = -\ell''(\theta) - \frac{\partial^2}{\partial\theta\, \partial\theta^T} \log p(\theta)$$

Here the negative Hessian of the log-likelihood is called the **observed (Fisher) information (matrix)**, and we denote it by $J(\theta)$,

$$J(\theta) = -\ell''(\theta) = -\frac{\partial^2}{\partial\theta\, \partial\theta^T} \log p(y \mid \theta). \tag{6.9}$$

The negative Hessian of the log-posterior equals the sum of the observed information and the negative Hessian of the log-prior.

If the sample size is large, then the likelihood dominates the prior in the sense that the likelihood is highly peaked while the prior is relatively flat in the region where the posterior density is appreciable. In large samples the mode of the log-posterior $\hat{\theta}$ and the mode of the log-likelihood (the maximum likelihood estimator, MLE) $\hat{\theta}_{\mathrm{MLE}}$ are approximately equal, and also the Hessian matrix of the log-posterior is approximately the same as the Hessian of the log-likelihood. Combining these two approximations, we get

$$\hat{\theta} \approx \hat{\theta}_{\mathrm{MLE}}, \qquad -L''(\hat{\theta}) \approx J(\hat{\theta}_{\mathrm{MLE}}).$$

When we plug these approximations in the normal approximation (6.8), we see that in large samples the posterior is approximately normal with mean equal to the MLE and covariance matrix given by the inverse of the observed information,

$$p(\theta \mid y) \approx N\left(\theta \mid \hat{\theta}_{\mathrm{MLE}}, [J(\hat{\theta}_{\mathrm{MLE}})]^{-1}\right). \tag{6.10}$$

This approximation should be compared with the well-known frequentist asymptotic distribution results for the maximum likelihood estimator. Loosely, these results can be summarized so that the sampling distribution of the maximum likelihood estimator is asymptotically normal with mean equal to the MLE and covariance matrix equal to the inverse of the observed information. In order to write this approximation as a formula, we need to indicate the dependence of the maximum likelihood estimator on the data as follows,

$$\hat{\theta}_{\mathrm{MLE}}(Y) \overset{\mathrm{d}}{\approx} N\left(\hat{\theta}_{\mathrm{MLE}}(y), [J(\hat{\theta}_{\mathrm{MLE}}(y))]^{-1}\right). \tag{6.11}$$

Here Y is a random vector from the sampling distribution of the data, and so $\hat{\theta}_{\mathrm{MLE}}(Y)$ is the maximum likelihood estimator considered as a random variable (or random vector). In contrast, $\hat{\theta}_{\mathrm{MLE}}(y)$ is the maximum likelihood estimate calculated from the observed data y.

Comparing equations (6.10) and (6.11) we see that for large samples the posterior distribution can be approximated using the same formulas that (frequentist) statisticians use for the maximum likelihood estimator. In large samples the influence of the prior vanishes, and then one does need to spend much energy on formulating the prior distribution so that it would reflect all available prior information. However, in small samples careful formulation of the prior is important.

6.3 Posterior expectations using Laplace approximation

Laplace showed in the 1770's how one can form approximations to integrals of highly peaked positive functions by integrating analytically a suitable normal approximation. We will now apply this idea to build approximations to posterior expectations. We assume that the posterior density is highly peaked while the function k, whose posterior expectation we seek is relatively flat. The posterior

density is typically known only in the unnormalized form $q(\theta \mid y)$, and then

$$E[k(\Theta) \mid Y = y] = \frac{\int k(\theta)\, q(\theta \mid y)\, \mathrm{d}\theta}{\int q(\theta \mid y)\, \mathrm{d}\theta}. \tag{6.12}$$

Tierney and Kadane [4] approximated separately the numerator and the denominator of eq. (6.12) using Laplace's method, and analyzed the resulting error.

To introduce the idea of Laplace's approximation (or Laplace's method), consider a highly peaked function $L(\theta)$ of a scalar variable θ such that $L(\theta)$ has a unique mode (i.e., a maximum) at $\hat{\theta}$. Suppose that $g(\theta)$ is a function, which varies slowly. We seek an approximation to the integral

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta. \tag{6.13}$$

Heuristically, the integrand is negligible when we go far away from $\hat{\theta}$, and so we should be able to approximate the integral I by a simpler integral, where we take into account only the local behavior of $L(\theta)$ around its mode. To this end, we first approximate $L(\theta)$ by its second degree Taylor polynomial centered at the mode $\hat{\theta}$,

$$L(\theta) \approx L(\hat{\theta}) + 0 \cdot (\theta - \hat{\theta}) + \frac{1}{2} L''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Since $g(\theta)$ is slowly varying, we may approximate the integrand as follows

$$g(\theta)\, \mathrm{e}^{L(\theta)} \approx g(\hat{\theta})\, \exp\left(L(\hat{\theta}) - \frac{1}{2} Q(\theta - \hat{\theta})^2) \right),$$

where

$$Q = -L''(\hat{\theta}).$$

For the following, we must assume that $L''(\hat{\theta}) < 0$. Integrating the approximation, we obtain

$$\begin{aligned} I &\approx \int g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}\, \exp(-\frac{1}{2} Q(\theta - \hat{\theta})^2)\, \mathrm{d}\theta \\ &= \frac{\sqrt{2\pi}}{\sqrt{Q}}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})} \end{aligned} \tag{6.14}$$

This is the univariate case of Laplace's approximation. (Actually, it is just the leading term in a Laplace expansion, which is an asymptotic expansion for the integral.)

To handle the multivariate result, we use the normalizing constant of the $N_d(\mu, Q^{-1})$ distribution to evaluate the integral

$$\int \exp\left(-\frac{1}{2} (x - \mu)^T Q (x - \mu) \right)\, \mathrm{d}x = \frac{(2\pi)^{d/2}}{\sqrt{\det Q}}. \tag{6.15}$$

This result is valid for any symmetric and positive definite $d \times d$ matrix Q. Integrating the multivariate second degree approximation of $g(\theta)\, \exp(L(\theta))$, we obtain

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta \approx \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}, \tag{6.16}$$

where d is the dimensionality of θ, and Q is the negative Hessian of L evaluated at the mode,

$$Q = -L''(\hat{\theta}),$$

and we must assume that the $d \times d$ matrix Q is positively definite.

Using these tools, we can approximate the posterior expectation of $k(\theta)$ (see (6.12)) in several different ways. One idea is to approximate the numerator by choosing

$$g(\theta) = k(\theta), \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

in eq. (6.16), and then to approximate the denominator by choosing

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y).$$

These choices yield the approximation

$$E[h(\Theta) \mid Y = y] \approx \frac{\frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, k(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}}{\frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, \mathrm{e}^{L(\hat{\theta})}} = h(\hat{\theta}), \tag{6.17}$$

where

$$\hat{\theta} = \arg\max L(\theta), \qquad Q = -L''(\hat{\theta}).$$

Here we need a single maximization, and do not need to evaluate the Hessian at all.

A less obvious approach is to choose

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = k(\theta)\, q(\theta \mid y)$$

to approximate the numerator, and

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

to approximate the denominator. Here we need to assume that h is a positive function, i.e., $h > 0$. The resulting approximation is

$$E[h(\Theta) \mid Y = y] \approx \left(\frac{\det(Q)}{\det(Q^*)}\right)^{1/2} \frac{k(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)}, \tag{6.18}$$

where

$$\hat{\theta}^* = \arg\max[k(\theta)\, q(\theta \mid y)], \qquad \hat{\theta} = \arg\max q(\theta \mid y).$$

and Q^* and Q are the negative Hessians

$$Q^* = -L^{*''}(\hat{\theta}^*), \qquad Q = -L''(\hat{\theta}),$$

where

$$L^*(\theta) = \log(k(\theta)\, q(\theta \mid y)), \qquad L(\theta) = \log q(\theta \mid y).$$

We need two separate maximizations and need to evaluate two Hessians for this approximation.

Tierney and Kadane analyzed the errors committed in these approximations in the situation, where we have n (conditionally) i.i.d. observations, and the

sample size n grows. The first approximation (6.17) has relative error of order $O(n^{-1})$, while the second approximation (6.18) has relative error of order $O(n^{-2})$. That is,

$$E[k(\Theta) \mid Y = y] = k(\hat{\theta}) \left(1 + O(n^{-1})\right)$$

and

$$E[k(\Theta) \mid Y = y] = \left(\frac{\det(Q)}{\det(Q^*)}\right)^{1/2} \frac{k(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)} \left(1 + O(n^{-2})\right).$$

Hence the second approximation is much more accurate (at least asymptotically).

6.4 Posterior marginals using Laplace approximation

Tierney and Kadane discuss also an approximation to the marginal posterior, when the parameter vector θ is composed of two vector components $\theta = (\phi, \psi)$. The form of the approximation is easy to derive, and was earlier discussed by Leonard [1]. However, Tierney and Kadane [4, Sec. 4] were the first to analyze the error in this Laplace approximation. We first derive the form of the approximation, and then make some comments on the error terms based on the discussion of Tierney and Kadane.

Let $q(\phi, \psi \mid y)$ be an unnormalized form of the posterior density, based on which we try to approximate the normalized marginal posterior $p(\phi \mid y)$. Let the dimensions of ϕ and ψ be d_1 and d_2, respectively. We have

$$p(\phi \mid y) = \int p(\phi, \psi \mid y)\, \mathrm{d}\psi = \int \exp(\log p(\phi, \psi \mid y))\, \mathrm{d}\psi,$$

where $p(\phi, \psi \mid y)$ is the normalized posterior. The main difference with approximating a posterior expectation is the fact, that now we are integrating only over the component(s) ψ of $\theta = (\phi, \psi)$.

Fix the value of ϕ for the moment. Let $\psi^*(\phi)$ be the maximizer of the function

$$\psi \mapsto \log p(\phi, \psi \mid y),$$

and let $Q(\phi)$ be the negative Hessian matrix of this function evaluated at $\psi = \psi^*(\phi)$. Notice that we can equally well calculate $\psi^*(\phi)$ and $Q(\phi)$ as the maximizer and the negative of the $d_2 \times d_2$ Hessian matrix of $\psi \mapsto \log q(\phi, \psi \mid y)$, respectively,

$$\psi^*(\phi) = \arg\max_{\psi} \left(\log q(\phi, \psi \mid y)\right) = \arg\max_{\psi} q(\phi, \psi \mid y) \tag{6.19}$$

$$Q(\phi) = -\left[\frac{\partial^2}{\partial\psi\, \partial\psi^T} \log q(\phi, \psi \mid y)\right]_{|\psi=\psi^*(\phi)}. \tag{6.20}$$

For fixed ϕ, we have the second degree Taylor approximation in ψ,

$$\log p(\phi, \psi \mid y) \approx \log p(\phi, \psi^*(\phi) \mid y) - \frac{1}{2}(\psi - \psi^*(\phi))^T Q(\phi)(\psi - \psi^*(\phi)), \tag{6.21}$$

and we assume that matrix $Q(\phi)$ is positive definite.

Next we integrate the exponential function of the approximation (6.21) with respect to ψ, with the result

$$p(\phi \mid y) \approx p(\phi, \psi^*(\phi) \mid y)\, (2\pi)^{d_2/2}\, (\det Q(\phi))^{-1/2}.$$

To evaluate this approximation, we need the normalizing constant of the unnormalized posterior $q(\phi, \psi \mid y)$, which we obtain by another Laplace approximation, and the end result is

$$p(\phi \mid y) \approx (2\pi)^{-d_1/2}\, q(\phi, \psi^*(\phi) \mid y) \sqrt{\frac{\det Q}{\det Q(\phi)}}, \tag{6.22}$$

where Q is negative of the $(d_1 + d_2) \times (d_1 + d_2)$ Hessian of the function

$$(\phi, \psi) \mapsto \log q(\phi, \psi \mid y)$$

evaluated at the MAP, the maximum point of the same function. However, it is often enough to approximate the functional form of the marginal posterior. When considered as a function of ϕ, we have, approximately,

$$p(\phi \mid y) \propto q(\phi, \psi^*(\phi) \mid y)\, (\det Q(\phi))^{-1/2}. \tag{6.23}$$

The unnormalized Laplace approximation (6.23) can be given another interpretation (see, e.g., [2, 3]). By the multiplication rule,

$$p(\phi \mid y) = \frac{p(\phi, \psi \mid y)}{p(\psi \mid \phi, y)} \propto \frac{q(\phi, \psi \mid y)}{p(\psi \mid \phi, y)}.$$

This result is valid for any choice of ψ. Let us now form a normal approximation for the denominator for a fixed value of ϕ, i.e.,

$$p(\psi \mid \phi, y) \approx N(\psi \mid \psi^*(\phi), Q(\phi)^{-1}).$$

However, this approximation is accurate only in the vicinity of the mode $\psi^*(\phi)$, so let us use it only at the mode. The end result is the following approximation,

$$\begin{aligned} p(\phi \mid y) &\propto \left[\frac{q(\phi, \psi \mid y)}{N(\psi \mid \psi^*(\phi), Q(\phi)^{-1})} \right]_{|\psi = \psi^*(\phi)} \\ &= (2\pi)^{d_2/2} \det(Q(\phi))^{-1/2}\, q(\phi, \psi^*(\phi) \mid y) \\ &\propto q(\phi, \psi^*(\phi) \mid y)\, (\det Q(\phi))^{-1/2}, \end{aligned}$$

which is the same as the unnormalized Laplace approximation (6.23) to the marginal posterior of ϕ.

Tierney and Kadane show that the relative error in the approximation (6.22) is of the order $O(n^{-1})$, when we have n (conditionally) i.i.d. observations, and that most of the error comes from approximating the normalizing constant. They argue that the approximation (6.23) captures the correct functional form of the marginal posterior with relative error $O(n^{-3/2})$ and recommend that one should therefore use the unnormalized approximation (6.23), which can then be normalized by numerical integration, if need be. For instance, if we

want to simulate from the approximate marginal posterior, then we can use the unnormalized approximation (6.23) directly, together with accept–reject, SIR or the grid-based simulation method of Sec. 6.1. See the articles by H. Rue and coworkers [2, 3] for imaginative applications of these ideas.

Another possibility for approximating the marginal posterior would be to build a normal approximation to the joint posterior, and then marginalize. However, a normal approximation to the marginal posterior would only give the correct result with absolute error of order $O(n^{-1/2})$, so the accuracies of both of the Laplace approximations are much better. Since the Laplace approximations yield good relative instead of absolute error, the Laplace approximations maintain good accuracy also in the tails of the densities. In contrast, the normal approximation is accurate only in the vicinity of the mode.

Example 6.2. Consider normal observations

$$[Y_i \mid \mu, \tau] \overset{\text{i.i.d.}}{\sim} N(\mu, \frac{1}{\tau}), \qquad i = 1, \ldots, n,$$

together with the non-conjugated prior

$$p(\mu, \tau) = p(\mu)\, p(\tau) = N(\mu \mid \mu_0, \frac{1}{\psi_0})\ \mathrm{Gam}(\tau \mid a_0, b_0).$$

The full conditional of μ is readily available,

$$p(\mu \mid \tau, y) = N(\mu \mid \mu_1, \frac{1}{\psi_1})$$

where

$$\psi_1 = \psi_0 + n\tau \qquad \psi_1\, \mu_1 = \psi_0\, \mu_0 + \tau \sum_{i=1}^{n} y_i$$

The mode of the full conditional $p(\mu \mid \tau, y)$ is

$$\mu^*(\tau) = \mu_1 = \frac{\psi_0\, \mu_0 + \tau \sum_{i=1}^{n} y_i}{\psi_0 + n\tau}.$$

We now use this knowledge to build a Laplace approximation to the marginal posterior of τ.

Since, as a function of μ,

$$p(\mu, \tau \mid y) \propto p(\mu \mid \tau, y),$$

$\mu^*(\tau)$ is also the mode of $p(\mu, \tau \mid y)$ for any τ. We also need the second derivative

$$\frac{\partial^2}{\partial \mu^2} \left(\log p(\mu, \tau \mid y)\right) = \frac{\partial^2}{\partial \mu^2} \left(\log p(\mu \mid \tau, y)\right) = -\psi_1,$$

for $\mu = \mu^*(\tau)$, but the derivative does not in this case depend on the value of μ at all. An unnormalized form of the Laplace approximation to the marginal posterior of τ is therefore

$$p(\tau \mid y) \propto \frac{q(\mu^*(\tau), \tau \mid y)}{\sqrt{\psi_1}}, \quad \text{where} \quad q(\mu, \tau \mid y) = p(y \mid \mu, \tau)\, p(\mu)\, p(\tau).$$

In this toy example, the Laplace approximation (6.23) for the functional form of the marginal posterior $p(\tau \mid \mu)$ is exact, since by the multiplication rule,

$$p(\tau \mid y) = \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)}$$

for any choice of μ, in particular for $\mu = \mu^*(\tau)$. Here the numerator is known only in an unnormalized form.

Figure 6.2 (a) illustrates the result using data $y = (-1.4, -1.6, -2.4, 0.7, 0.6)$ and hyperparameters $\mu_0 = 0$, $\psi_0 = 0.5$, $a_0 = 1$, $b_0 = 0.1$. The unnormalized (approximate) marginal posterior has been drawn using the grid method of Sec. 6.1. Figure 6.2 (b) shows an i.i.d. sample drawn from the approximate posterior

$$\tilde{p}(\tau \mid y)\, p(\mu \mid \tau, y),$$

where $\tilde{p}(\tau \mid y)$ is a histogram approximation to the true marginal posterior $p(\tau \mid y)$, which has been sampled using the grid method.

△

Bibliography

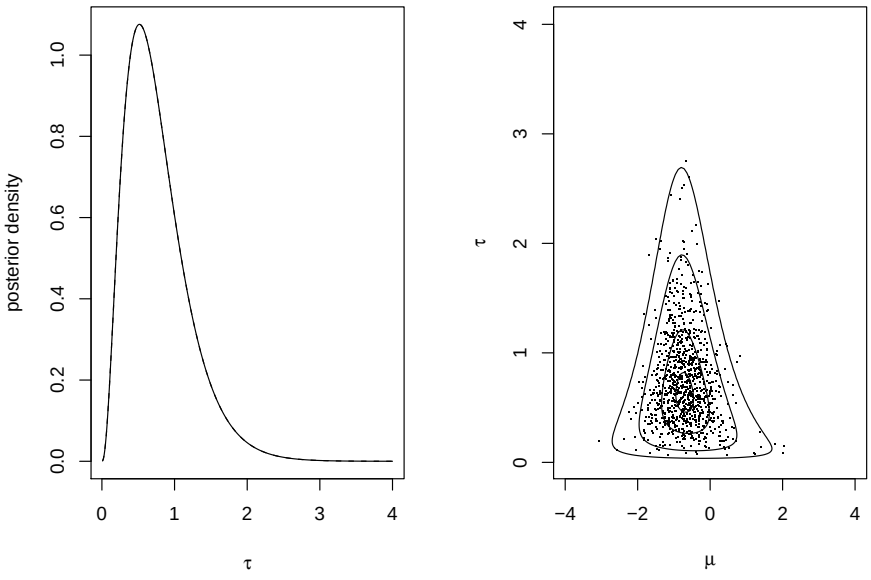
[1] Tom Leonard. A simple predictive density function: Comment. *Journal of the American Statistical Association*, 77:657–658, 1982.

[2] H. Rue and S. Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192, 2007.

[3] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Lapalce approximations. *Journal of the Royal Statistical Society: Series B*, 2009. to appear.

[4] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

Figure 6.2: (a) Marginal posterior density of τ and (b) a sample drawn from the approximate joint posterior together with contours of the true joint posterior density.



(a) (b)

Chapter 7

MCMC algorithms

7.1 Introduction

In a complicated Bayesian statistical model it may be very difficult to analyze the mathematical form of the posterior and it may be very difficult to draw an i.i.d. sample from it. Fortunately, it is often easy to generate a correlated sample, which approximately comes from the posterior distribution. (In this context, the word *correlated* means *not independent*). However, we would very much prefer to have an i.i.d. sample from the posterior, instead. After one has available a sample, one can estimate posterior expectations and posterior quantiles using the same kind of techniques that are used with i.i.d. samples. This is the idea behind Markov chain Monte Carlo (MCMC) methods.

In this chapter we will introduce the basic MCMC sampling algorithms that are used in practical problems. The emphasis is on trying to understand what one needs to do in order to implement the algorithms. In Chapter 11 we will see why these algorithms work using certain concepts from the theory of Markov chains in a general state space.

There are available computer programs that can implement an MCMC simulation automatically. Perhaps the most famous such program is the BUGS system (Bayesian inference Using Gibbs Sampling), which has several concrete implementations, most notably WinBUGS and OpenBUGS. You can analyze most of the models of interest easily using BUGS. What the user of BUGS needs to do is to write the description of the model in a format that BUGS understands, read the data into the program, and then let the program do the simulation. Once the simulation has finished, one can let the program produce various summaries of the posterior. Using such a tool, it is simple to experiment with different priors and different likelihoods for the same data.

However, in this chapter the emphasis is on understanding how you can write your own MCMC programs. Why would this be of interest?

- If you have not used MCMC before, you get a better understanding of the methods if you try to implement (some of) them yourself.
- For some models, the automated tools fail. Sometimes you can, however, rather easily design and implement a MCMC sampler yourself, once you understand the basic principles. (In some cases, however, designing an efficient MCMC sampler can be an almost impossibly difficult task.)

- Sometimes you want to have more control over the sampling algorithm than is provided by the automated tools. In some cases implementation details can make a big difference to the efficiency of the method.

The most famous MCMC methods are the Metropolis–Hastings sampler and the Gibbs sampler. Where do these names come from?

- Nicholas (Nick) Metropolis (1915–1999) was an American mathematician, physicist and pioneer of computing, who was born in Greece. He published the Metropolis sampler in 1953 jointly with two husband-and-wife teams, namely A.W. and M.N. Rosenbluth and A.H. and E. Teller. At that time the theory of general state space Markov chains was largely unexplored. In spite of this, the authors managed to give a heuristic proof for the validity of the method.

- W. Keith Hastings (1930–) is a Canadian statistician, who published the Metropolis–Hastings sampler in 1970. It is a generalization of the Metropolis sampler. Hastings presented his algorithm using a discrete state space formalism, since the theory of general state space Markov chains was then known only to some specialists in probability theory. Hastings' article did not have a real impact on statisticians until much later.

- The name Gibbs sampler was introduced by the brothers S. and D. Geman in an article published in 1984. Related ideas were published also by other people at roughly the same time. The method is named after the American mathematician and physicist J. Willard Gibbs (1893–1903), who studied thermodynamics and statistical physics, but did not have anything to do with MCMC.

In the late 1980's and early 1990's there was an explosion in the number of studies, where people used MCMC methods in Bayesian inference. Now there was available enough computing power to apply the methods, and besides, the theory of general state space Markov chains had matured so that readable expositions of the theory were available.

Nowadays, many statisticians routinely use the concept of a Markov chain which evolves in a general state space. Unfortunately, their mathematical theory is still explained only in a handful of text books.

7.2 Basic ideas of MCMC

MCMC algorithms are based on the idea of a Markov chain which evolves in discrete time. A Markov chain is a stochastic process

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$$

Here $\theta^{(i)}$ (the state of the process at time i) is a RV whose values lie in a state space, which usually is a subset of some Euclidean space $\mathbb{R}^d$. The state space is the same for all times i. We write the time index as a superscript so that we can index the components $\theta^{(i)}$ using a subscript.

Markov chains have the following **Markov property**: the distribution of the next state $\theta^{(i+1)}$ depends on the history $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(i)}$ only through the

present state $\theta^{(i)}$. The Markov chains used in MCMC methods are **homogeneous**: the conditional distribution of $\theta^{(i+1)}$ given $\theta^{(i)}$ does not depend on the index i.

The following algorithm shows how one can simulate a Markov chain, in principle. Intuitively, a Markov chain is nothing else but the mathematical idealization of this simulation algorithm. (There are, howere, important Markov chains which are easier to simulate using some other structure for the simulation program.)

Algorithm 14: Computer scientist's definition of a homogeneous Markov chain.

```
1 Generate θ^(0) from a given initial distribution;
2 for i = 0, 1, 2, ... do
3     Generate a vector V^(i+1) of fresh random numbers from a suitable
      distribution;
4     θ^(i+1) ← h(θ^(i), V^(i+1)) for a suitable function h(·, ·);
5 end
```

Some (but not all) Markov chains have an **invariant distribution** (or a stationary distribution or equilibrium distribution), which can be defined as follows. If the initial state of the chain $\theta^{(0)}$ follows the invariant distribution, then also all the subsequent states $\theta^{(i)}$ follow it.

If a Markov chain has an invariant distribution, then (under certain regularity conditions) the distribution of the state $\theta^{(i)}$ converges to that invariant distribution (in a certain sense). Under certain regularity conditions, such a chain is **ergodic**, which ensures that an arithmetic average (or an ergodic average) of the form

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)})$$

converges, almost surely, to the corresponding expectation calculated under the invariant distribution as $N \to \infty$. That is, the ergodic theorem for Markov chains then states that the strong law of large numbers holds, i.e.,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \to E_f h(\Theta) = \int h(\theta) f(\theta)\,\mathrm{d}\theta, \tag{7.1}$$

where f is the density of the invariant distribution. This will then hold for all functions h for which the expectation $E_f h(\Theta)$ exists, so the convergence is as strong as in the strong law of large numbers for i.i.d. sequences. There are also more advanced forms of ergodicity (geometric ergodicity and uniform ergodicity), which a Markov chain may either have or not have.

Under still more conditions, Markov chains also satisfy a central limit theorem, which characterizes the speed of convergence in the ergodic theorem. The central limit theorem for Markov chains is of the form

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) - E_f h(\Theta)\right) \xrightarrow{\mathrm{d}} N(0, \sigma_h^2).$$

The speed of convergence is of the same order of N as in the central limit theorem for i.i.d. sequences. However, estimating the variance σ_h^2 in the central limit theorem is lot trickier than with i.i.d. sequences.

After this preparation, it is possible to explain the basic idea of MCMC methods. The idea is to set up an ergodic Markov chain which has the posterior distribution as its invariant distribution. Doing this is often surprisingly easy. Then one simulates values

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$$

of the chain. When t is sufficiently large, then $\theta^{(t)}$ and all the subsequent states $\theta^{(t+i)}, i \geq 1$ follow approximately the posterior distribution. The time required for the chain to approximately achieve its invariant distribution is called the **burn-in**. After the initial burn-in period has been discarded, the subsequent values

$$\theta^{(t)}, \theta^{(t+1)}, \theta^{(t+2)}, \ldots$$

can be treated as a dependent sample from the posterior distribution, and we can calculate posterior expectations, quantiles and other summaries of the posterior distribution based on this sample.

After the burn-in period we need to store the simulated values of the chain for later use. So, for a scalar parameter we need a vector to store the results, for a vector parameter we need a matrix to store the results and so on. To save space, one often decides to **thin** the sequences by keeping only every kth value of each sequence and by discarding the rest.

Setting up *some* MCMC algorithm for a given posterior is usually easy. However, the challenge is to find an MCMC algorithm which converges rapidly and then explores efficiently the whole support of the posterior distribution. Then one can get a reliable picture of the posterior distribution after stopping the simulation after a reasonable number of iterations.

In practice one may want to try several approaches for approximate posterior inference in order to become convinced that the posterior inferences obtained with MCMC are reliable. One can, e.g., study simplified forms of the statistical model (where analytical developments or maximum likelihood estimation or other asymptotic approximations to Bayesian estimation may be possible), simulate several chains which are initialized from different starting points and are possibly computed with different algorithms, and compute approximations to the posterior.

7.3 The Metropolis–Hastings algorithm

Now we consider a target distribution with density $\pi(\theta)$, which may be available only in an unnormalized form $\tilde{\pi}(\theta)$. Usually the target density is the posterior density of a Bayesian statistical model,

$$\pi(\theta) = p(\theta \mid y).$$

Actually we only need to know an unnormalized form of the posterior, which is given, e.g., in the form of prior times likelihood,

$$\tilde{\pi}(\theta) = p(\theta)\, p(y \mid \theta).$$

The density $\pi(\theta)$ may be a density in the generalized sense, so we may have a discrete distribution for some components of θ and a continuous distribution for others.

For the Metropolis–Hastings algorithm we need a proposal density $q(\theta' \mid \theta)$, from which we are able to simulate. (Some authors call the proposal density the jumping density or candidate generating density.) As a function of θ', the proposal density $q(\theta' \mid \theta)$ is a density on the parameter space for each value of θ. When the current state of the chain is $\theta = \theta^{(i)}$, we propose a value for the next state from the distribution with density

$$\theta' \mapsto q(\theta' \mid \theta)$$

The proposed value θ' is then accepted or rejected in the algorithm. If the proposal is accepted, then the next state $\theta^{(i+1)}$ is taken to be θ', but otherwise the chain stays in the same state, i.e., $\theta^{(i+1)}$ is assigned the current state $\theta^{(i)}$.

The acceptance condition has to be selected carefully so that we get the target distribution as the invariant distribution of the chain. The usual procedure works as follows. We calculate the value of the Metropolis–Hastings ratio (M–H ratio)

$$r = r(\theta', \theta) = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta),} \tag{7.2}$$

where $\theta = \theta^{(i)}$ is the current state and θ' is the proposed state. Then we generate a value u from the standard uniform $\mathrm{Uni}(0, 1)$. If $u < r$, then we accept the proposal and otherwise reject it. For the analysis of the algorithm, it is essential to notice that the probability of accepting the proposed θ', when the current state is θ, is given by

$$\Pr(\text{proposed value is accepted} \mid \theta^{(i)} = \theta, \theta') = \min(1, r(\theta', \theta)). \tag{7.3}$$

We need here the minimum of one and the M–H ratio, since the M–H ratio may very well be greater than one.

Some explanations are in order.

- The denominator of the M–H ratio (7.2) is the joint density of the proposal θ' and the current state θ, when the current state already follows the posterior.
- The numerator is of the same form as the denominator, but θ and θ' have exchanged places.
- If $\pi(\theta^{(0)}) > 0$, then the denominator of the M–H ratio is always strictly positive during the algorithm. When $i = 0$ this follows from the observation that $q(\theta' \mid \theta^{(0)})$ has to be positive, since θ' is generated from that density. Also $\pi(\theta^{(1)})$ has to be positive, thanks to the form of the acceptance test. The rest follows by induction.
- We do not need to know the normalizing constant of the target distribution, since it cancels in the M–H ratio,

$$r = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} = \frac{\tilde{\pi}(\theta')\, q(\theta \mid \theta')}{\tilde{\pi}(\theta)\, q(\theta' \mid \theta)} \tag{7.4}$$

- If the target density is a posterior distribution, then the M–H ratio is given by

$$r = \frac{f_{Y|\Theta}(y \mid \theta')\, f_\Theta(\theta')\, q(\theta \mid \theta')}{f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta)\, q(\theta' \mid \theta)}. \tag{7.5}$$

- Once you know what the notation is supposed to mean, you can use an abbreviated notation for the M–H ratio, such as

$$r = \frac{p(\theta' \mid y)\, q(\theta \mid \theta')}{p(\theta \mid y)\, q(\theta' \mid \theta)}.$$

  Here, e.g., $p(\theta' \mid y)$ is the value of the posterior density evaluated at the proposal θ'.

An explanation of why the target distribution is the invariant distribution of the resulting Markov chain will be given in Chapter 11. Then it will become clear, that other formulas in place of eq. (7.2) would work, too. However, the formula (7.2) is known to be optimal (in a certain sense), and therefore it is the one that is used in practice.

In the Metropolis–Hastings algorithm the proposal density can be selected otherwise quite freely, but we must be sure that we can reach (with positive probability) any reasonably possible region in the parameter space starting from any initial state $\theta^{(0)}$ with a finite number of steps. This property is called **irreducibility** of the Markov chain.

Algorithm 15: The Metropolis–Hastings algorithm.

Input: An initial value $\theta^{(0)}$ such that $\tilde{\pi}(\theta^{(0)}) > 0$ and the number of iterations N.

Result: Values simulated from a Markov chain which has as its invariant distribution the distribution corresponding to the unnormalized density $\tilde{\pi}(\theta)$.

1 **for** $i = 0, 1, 2, \ldots, N$ **do**

2 $\quad \theta \leftarrow \theta^{(i)}$;

3 $\quad$ Generate θ' from $q(\cdot \mid \theta)$ and u from $\mathrm{Uni}(0, 1)$;

4 $\quad$ Calculate the M–H ratio

$$r = \frac{\tilde{\pi}(\theta')\, q(\theta \mid \theta')}{\tilde{\pi}(\theta)\, q(\theta' \mid \theta)}$$

5 $\quad$ Set

$$\theta^{(i+1)} \leftarrow \begin{cases} \theta', & \text{if } u < r \\ \theta, & \text{otherwise.} \end{cases}$$

6 **end**

Algorithm 15 sums up the Metropolis–Hastings algorithm. When implementing the algorithm, one easily comes across problems, which arise because of underflow or overflow in the calculation of the M–H ratio r. Most of such problems can be cured by calculating with logarithms. E.g., when the target distribution is a posterior distribution, then one should first calculate $s = \log r$ by

$$\begin{aligned} s = {} & \log(f_{Y|\Theta}(y \mid \theta')) - \log(f_{Y|\Theta}(y \mid \theta)) \\ & + \log(f_\Theta(\theta')) - \log(f_\Theta(\theta)) + \log(q(\theta \mid \theta')) - \log(q(\theta' \mid \theta)) \end{aligned}$$

and only then calculate $r = \exp(s)$. Additionally, one might want cancel common factors from r before calculating its logarithm.

Implementing some Metropolis–Hastings algorithm for any given Bayesian statistical model is usually straightforward. However, finding a proposal distribution which allows the chain to converge quickly to the target distribution and allows it to explore the parameter space efficiently may be challenging.

7.4 Concrete Metropolis–Hastings algorithms

In the Metropolis–Hastings algorithm, the proposal θ' is in practice generated by a piece of code, which can use the current state $\theta^{(i)}$, freshly generated random numbers from any distribution and arbitrary arithmetic operations. We must be able to calculate the (correctly normalized) density of the proposal θ', when the current state is equal to θ. This is then $q(\theta' \mid \theta)$, which we must be able to evaluate. Or at least we must be able to calculate the value of the ratio

$$q(\theta \mid \theta')/q(\theta' \mid \theta).$$

Different choices for the proposal density correspond to different choices for the needed piece of code. The resulting Metropolis–Hastings algorithms are named after the properties of the proposal distribution. We next look at some widely-used examples.

7.4.1 The independent Metropolis–Hastings algorithm

In the independent M–H algorithm (other common names: independence chain independence sampler), the proposal density is a fixed density, say $s(\theta')$, which does not depend on the value of the current state. In the corresponding piece of code, we only need to generate the value θ' from the proposal distribution.

If the proposal distribution happens to be the target distribution, then every proposal will be accepted, and as a result we will get an i.i.d. sample from the target distribution.

In order to to sample the target distribution properly with the independent M–H algorithm, the proposal density s must be positive everywhere, where the target density is positive. If there exist a majorizing constant M, such that

$$\pi(\theta) \leq Ms(\theta) \qquad \forall\theta,$$

then the resulting chain can be shown to have good ergodic properties, but if this condition fails, then the convergence properties of the chain can be bad. (In the independent M–H algorithm one does not need to know the value of M.) This implies that the proposal density should be such that the accept–reject method or importance sampling using that proposal distribution would be possible, too.

In particular, the tails of the proposal density s should be at least as heavy as the tails of the target density. Finding such proposal densities may be difficult in high-dimensional problems. A natural choice would be a multivariate t distribution whose shape is chosen to match the shape of the posterior density. One should choose a low value (e.g. $\nu = 4$) for the degrees of freedom parameter in order to ensure heavy tails, and then one could choose the center μ of the multivariate t distribution $t(\nu, \mu, \Sigma)$ to be equal to the posterior mode and the

dispersion parameter Σ to be equal to the covariance matrix of an approximating normal distribution. Other choices for the center and dispersion matrix are possible, too. E.g., one could choose μ to be equal to the estimated posterior mean and Σ equal to the posterior covariance matrix.

7.4.2 Symmetric proposal distribution

If the proposal density is symmetric in that

$$q(\theta' \mid \theta) = q(\theta \mid \theta'), \qquad \forall \theta, \theta',$$

then the proposal density cancels from the M–H ratio,

$$r = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} = \frac{\pi(\theta')}{\pi(\theta)}.$$

This the sampling method that was originally proposed by Metropolis. Proposals leading to a higher value for the target density are automatically accepted, and other proposals may be accepted or rejected. Later Hastings generalized the method for non-symmetric proposal densities.

7.4.3 Random walk Metropolis–Hastings

Suppose that g is a density on the parameter space an that we calculate the proposal as follows,

$$\text{generate } w \text{ from density } g \text{ and set } \theta' \leftarrow \theta + w.$$

Then the proposal density is

$$q(\theta' \mid \theta) = g(\theta' - \theta).$$

This kind of a proposal is called a random walk proposal. If the density g is symmetric, i.e.,

$$g(-w) = g(w) \quad \forall w,$$

then the proposal density $q(\theta' \mid \theta)$ is also symmetric, and thus cancels from the M–H ratio. In the case of a symmetric random walk proposal, one often speaks of the random walk Metropolis (RWM) algorithm.

Actually, a random walk is a stochastic process of the form $X_{t+1} = X_t + w_t$, where the random variables w_t are i.i.d. Notice that the stochastic process produced by the random walk M–H algorithm is **not** a random walk, since the proposals can either be accepted or rejected.

The symmetric random walk Metropolis (RWM) algorithm is one of the most commonly used forms of the Metropolis–Hastings method. The most commonly used forms for g are the multivariate normal or multivariate Student's t density centered at the origin. This is, of course, appropriate only for continuous posterior distributions.

In the preceeding discussion we have implicitely assumed that the parameter space is *unconstrained*, i.e., that it is equal to $\mathbb{R}^d$ for some dimensionality d. However, the parameter spaces of many important statistical models are constrained. The most typical constraints are positivity constraints (e.g., for

a variance parameter) or the constraint that a probability parameter should lie in the interval $(0, 1)$. The RWM algorithm can be used also in constrained parameter spaces using one of the following approaches.

The first approach is to reparametrize the constrained parameters so that they all become unconstrained. For example, one can take the logarithm of a positive parameter and the logit of a probability parameter as the new parameters. In this case one should take the Jacobian of the transformation into account, as is explained in Section 7.4.4. The second approach uses the original constrained parametrization in the following manner. Whenever the proposed new parameter vector θ' lies outside the parameter space, the proposal is immediately rejected, i.e., the chain then stays at its current state and the iteration counter is increased. The correctness of the second approach can be justified by embedding the original model in an unconstrained model, which has the same likelihood and prior as the original model when θ is inside the original parameter space, but where the prior $p(\theta)$ is zero and the sampling density $y \mapsto p(y \mid \theta)$ is defined to be some arbitrary density when θ is outside the original parameter space. The posterior distribution for the unconstrained model is the same as for the original model, and the RWM algorithm for the unconstrained model works as is described above.

In a RWM algorithm one often selects the covariance matrix of the proposal distribution as

$$a\,C,$$

where C is an approximation to the covariance matrix of the target distribution (in Bayesian inference C is an approximation to the posterior covariance matrix) and the scalar a is a tuning constant which should be calibrated carefully. These kind of proposal distributions work well when the posterior distribution is approximately normal. One sometimes needs to reparametrize the model in order to make this approach work better.

The optimal value of a and the corresponding optimal acceptance rate has been derived theoretically, when the target density is a multivariate normal $N_d(\mu, C)$ and the random walk proposal is $N_d(0, aC)$, see [14]. The scaling constant a should be about $(2.38)^2/d$ when d is large. The corresponding acceptance rate (the number of accepted proposals divided by the total number of proposals) is from around 0.2 (for high-dimensional problems) to around 0.4 (in dimensions one or two). While these results have been derived using the very restrictive assumption that the target density is a multivariate normal, the results anyhow give rough guidelines for calibrating a in a practical problem.

How and why should one try to control the acceptance rate in the random walk M–H algorithm? If the acceptance rate is too low, then the chain is not able to move, and the proposed updating steps are likely to be too large. In this case one could try a smaller value for a. However, a high acceptance rate may also signal a problem, since then the updating steps may be too small. This may lead to the situation where the chain explores only a small portion of the parameter space. In this case one should try a larger value for a. From the convergence point of view, too high acceptance rate is a bigger problem. A low acceptance rate is a problem only from the computing time point of view.

In order to calibrate the random walk M–H algorithm, one needs an estimate of its acceptance rate. A simple-minded approach is just to keep track of the number of accepted proposals. A better approach is to calculate the average of

the acceptance probabilities,

$$\frac{1}{N}\sum_{i=1}^{N}\min(1, r_i),$$

where r_i is the M–H ratio in the ith iteration.

In practice, one can try to tune a iteratively, until the acceptance rate is acceptable. The tuning iterations are discarded, and the MCMC sample on which the inference is based is calculated using the fixed proposal distribution, whose scale a is the selected value. Fixing the proposal distribution is necessary, since the theory of the Metropolis–Hastings algorithm requires a homogeneous Markov chain, i.e., a proposal density $q(\theta' \mid \theta)$ which does not depend on the iteration index.

Recently, several researchers have developed adaptive MCMC algorithms, where the proposal distribution is allowed to change all the time during the iterations, see [1, 15] for reviews. Be warned that the design of valid adpative MCMC algorithms is subtle and that their analysis requires tools which are more difficult than the general state space Markov chain theory briefly touched upon in Chapter 11.

Example 7.1. Let us try the random walk chain for the target distribution $N(0, 1)$ by generating the increment from the normal distribution $N(0, \sigma^2)$ using the following values for the variance: a) $\sigma^2 = 4$ b) $\sigma^2 = 0.1$ c) $\sigma^2 = 40$. In situation a) the chain is initialized far away in the tails of the target distribution, but nevertheless it quickly finds its way to the main portion of the target distribution and then explores it efficiently. Such a chain is said to **mix** well. In situations b) and c) the chains are initialized at the center of the target distribution, but the chains mix less quickly. In situation b) the step length is too small, but almost all proposals get accepted. In situation c) the algorithm proposes too large steps, almost all of which get rejected. Figure 7.1 presents trace plots (or time series plots) of the chain in the three situations.

△

7.4.4 Reparametrization

Suppose that the posterior distribution of interest is a continuous distribution and that we have implemented functions for calculating the log-prior and the log-likelihood in terms of the parameter θ. Now we want to consider a diffeomorphic reparametrization

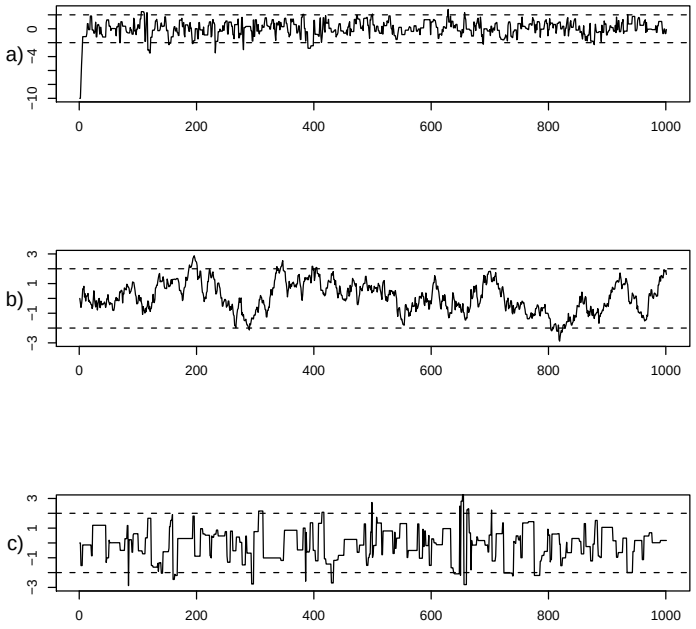
$$\phi = g(\theta) \quad \Leftrightarrow \quad \theta = h(\phi).$$

Typical reparametrizations one might consider are taking the logarithm of a positive parameter or calculating the logit function of a parameter constrained to the interval $(0, 1)$. What needs to be done in order to implement the Metropolis–Hastings algorithm for the new parameter vector ϕ?

First of all, we need a proposal density $q(\phi' \mid \phi)$ and the corresponding code. We also need to work out how to compute one of the Jacobians

$$J_h(\phi) = \frac{\partial\theta}{\partial\phi} \quad \text{or} \quad J_g(\theta) = \frac{\partial\phi}{\partial\theta}.$$

Figure 7.1: Trace plots of the random walk chain using the three different proposal distributions.



In ϕ-space the target density is given by the change of variables formula

$$f_{\Phi|Y}(\phi \mid y) = f_{\Theta|Y}(\theta \mid y) \left| \frac{\partial \theta}{\partial \phi} \right| = f_{\Theta|Y}(\theta \mid y) \, |J_h(\phi)| \,,$$

where $\theta = h(\phi)$.

The M–H ratio, when we propose ϕ' and the current value is ϕ, is given by

$$\begin{aligned} r &= \frac{f_{\Phi|Y}(\phi' \mid y)\, q(\phi \mid \phi')}{f_{\Phi|Y}(\phi \mid y)\, q(\phi' \mid \phi)} \\ &= \frac{f_{\Theta|Y}(\theta' \mid y)\, |J_h(\phi')|\, q(\phi \mid \phi')}{f_{\Theta|Y}(\theta \mid y)\, |J_h(\phi)|\, q(\phi' \mid \phi)} \\ &= \frac{f_{Y|\Theta}(y \mid \theta')\, f_\Theta(\theta')\, q(\phi \mid \phi')}{f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta)\, q(\phi' \mid \phi)} \, \frac{|J_h(\phi')|}{|J_h(\phi)|} \end{aligned}$$

here $\theta' = h(\phi')$ and $\theta = h(\phi)$. Sometimes it is more convenient to work with the Jacobian J_g, but this is easy, since

$$J_g(\theta) = \frac{1}{J_h(\phi)}.$$

Above we viewed the Jacobians as arising from expressing the target density using the new ϕ parametrization instead of the old θ parametrization. An alternative interpretation is that we should express the proposal density in θ space instead of ϕ space and then use the ordinary formula for M–H ratio. Both viewpoints yield the same formulas.

In order to calculate the logarithm of the M–H ratio, we need to do the following.

- Calculate the θ and θ' values corresponding to the current ϕ and proposed ϕ' values.
- Calculate the log-likelihood and log-prior using the values θ and θ'.
- Calculate the logarithm s of the M–H ratio as

$$\begin{aligned} s = \log(f_{Y|\Theta}(y \mid \theta')) - \log(f_{Y|\Theta}(y \mid \theta)) \\ + \log(f_\Theta(\theta')) - \log(f_\Theta(\theta)) + \log(q(\phi \mid \phi')) - \log(q(\phi' \mid \phi)) \\ + \log(|J_h(\phi')|) - \log(|J_h(\phi)|). \end{aligned}$$

  Finally, calculate $r = \exp(s)$.

- The difference of the logarithms of the absolute Jacobians can be calculated either on the ϕ scale or on the θ scale by using the identity

$$\log(|J_h(\phi')|) - \log(|J_h(\phi)|) = \log(|J_g(\theta)|) - \log(|J_g(\theta')|).$$

7.4.5 Langevin proposals

Unlike a random walk, the Langevin proposals introduce a drift which moves the chain towards the modes of the posterior distribution. When the current state is θ, the proposal θ' is generated with the rule

$$\theta' = \theta + \frac{\sigma^2}{2} \nabla(\log \pi(\theta)) + \sigma \, \epsilon, \qquad \epsilon \sim N_p(0, I).$$

Here $\sigma > 0$ is a tuning parameter and

$$\nabla(\log \pi(\theta)) = \nabla(\log \tilde{\pi}(\theta))$$

is the gradient of the logarithm of the (unnormalized) posterior density. The proposal distribution is motivated by a stochastic differential equation, which has π as its stationary distribution.

This proposal is then accepted or rejected using the ordinary Metropolis–Hastings rule, where the proposal density is

$$q(\theta' \mid \theta) = N_p(\theta' \mid \theta + \frac{\sigma^2}{2}\nabla(\log \pi(\theta)),\ \sigma^2 I).$$

7.4.6 State-dependent mixing of proposal distributions

Let θ be the current state of the chain. Suppose that the proposal θ' is drawn from a proposal density, which is selected randomly from a list of alternatives

$$q(\theta' \mid \theta, j), \qquad j = 1, \ldots K,$$

What is more, the selection probabilities may depend on the current state. One valid form for the step of an MCMC algorithm is then the following.

- Draw j from the pmf $\beta(\cdot \mid \theta), j = 1, \ldots, K$.
- Draw θ' from the density $q(\theta' \mid \theta, j)$ which corresponds to the selected j.
- Accept the proposed value θ' as the new state, if $U < r$, where $U \sim \mathrm{Uni}(0, 1)$, and
$$r = \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}. \tag{7.6}$$
Otherwise the chain stays at θ.

This formula (7.6) for the M–H ratio r is contained in Green's article [7], which introduced the reversible jump MCMC method. The algorithm could be called the Metropolis–Hastings–Green algorithm.

The lecturer does know any trick for deriving formula (7.6) from the M–H ratio of the ordinary M–H algorithm. The beauty of formula (7.6) lies in the fact that one only needs to evaluate $q(\theta' \mid \theta, j)$ and $q(\theta \mid \theta', j)$ for the proposal density which was selected. A straightforward application of the M–H algorithm would require one to evaluate these densities for all of the K possibilities.

If the selection probabilities $\beta(j \mid \theta)$ do not actually depend on θ, then they cancel from the M–H ratio. In this case (7.6) is easily derived from the ordinary M–H algorithm by augmenting the state of the chain to include the selected proposal mechanism.

7.5 Gibbs sampler

One of the best known ways of setting up an MCMC algorithm is Gibbs sampling, which is now discussed supposing that the target distribution is a posterior distribution. However, the method can be applied to any target distribution, when the full conditional distributions of the target distribution are available.

Suppose that the parameter vector has been divided into components

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d),$$

where θ_j may but need not be a scalar. Suppose also that the posterior full conditional distributions of each of the components are available in the sense that we know how to simulate them. This is the case when the statistical model exhibits conditional conjugacy with respect to all of the components θ_j. Then the basic idea behind Gibbs sampling is that we simulate successively each component θ_j from its (posterior) full conditional distribution. It is convenient to use the abbreviation θ_{-j} for the vector, which contains all the other components of θ but θ_j, i.e.

$$\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d). \tag{7.7}$$

Then the posterior full conditional of θ_j is

$$p(\theta_j \mid \theta_{-j}, y) = f_{\Theta_j|\Theta_{-j},Y}(\theta_j \mid \theta_{-j}, y). \tag{7.8}$$

A convenient shorthand notation for the posterior full conditional is

$$p(\theta_j \mid \cdot),$$

where the dot denotes all the other random variables except θ_j.

The most common form of the Gibbs sampler is the systematic scan Gibbs sampler, where the components are updated in a fixed cyclic order. It is also possible to select at random which component to update next. In that case one has the random scan Gibbs sampler.

Algorithm 16 presents the systematic scan Gibbs sampler, when we update the components using the order $1, 2, \ldots, d$. In the algorithm i is the time index of the Markov chain. One needs d updates to get from $\theta^{(i)}$ to $\theta^{(i+1)}$. To generate the j'th component, $\theta_j^{(i+1)}$, one uses the most recent values for the other components, some of which have already been updated. I.e., when the value for $\theta_j^{(i+1)}$ is generated, it is generated from the corresponding full conditional using the following values for the other components,

$$\theta_{-j}^{\text{cur}} = (\theta_1^{(i+1)}, \ldots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \ldots, \theta_d^{(i)}).$$

Usually the updating steps for the components of θ are so heterogeneous, that the inner loop is written out in full. E.g., in the case of three components, $\theta = (\phi, \psi, \tau)$, the actual implementation would probably look like the following algorithm 17. This algorithm also demonstrates, how one can write the algorithm using the abbreviated notation for conditional densities.

Algorithm 18 presents the random scan Gibbs sampler. Now one time step of the Markov chain requires only one update of a randomly selected component. In the random scan version, one can have different probabilities for updating the different components of θ, and this freedom can be useful for some statistical models.

If the statistical model exhibits conditional conjugacy with respect to all the components of θ, then the Gibbs sampler is easy to implement and is the method of choice for many statisticians. One only needs random number generators for all the posterior full conditionals, and these are easily available for the standard distributions. An appealing feature of the method is the fact that one does not

Algorithm 16: Systematic scan Gibbs sampler.

Input: An initial value $\theta^{(0)}$ such that $f_{\Theta|Y}(\theta^{(0)} \mid y) > 0$ and the number of iterations N.

Result: Values simulated from a Markov chain which has the posterior distribution as its invariant distribution.

$\theta^{\text{cur}} \leftarrow \theta^{(0)}$
for $i = 0, 1, \ldots, N$ **do**
  for $j = 1, \ldots, d$ **do**
    draw a new value for the jth component θ_j^{cur} of θ^{cur} from the posterior full conditional $f_{\Theta_j|\Theta_{-j},Y}(\theta_j \mid \theta_{-j}^{\text{cur}}, y)$
  end
  $\theta^{(i+1)} \leftarrow \theta^{\text{cur}}$
end

Algorithm 17: Systematic scan Gibbs sampler for three components $\theta = (\phi, \psi, \tau)$ given initial values for all the components except the one that gets updated the first.

$\psi^{\text{cur}} \leftarrow \psi_0$; $\tau^{\text{cur}} \leftarrow \tau_0$;
for $i = 0, 1, \ldots, N$ **do**
  draw ϕ^{cur} from $p(\phi \mid \psi = \psi^{\text{cur}}, \tau = \tau^{\text{cur}}, y)$;
  draw ψ^{cur} from $p(\psi \mid \phi = \phi^{\text{cur}}, \tau = \tau^{\text{cur}}, y)$;
  draw τ^{cur} from $p(\tau \mid \phi = \phi^{\text{cur}}, \psi = \psi^{\text{cur}}, y)$;
  $\phi_{i+1} \leftarrow \phi^{\text{cur}}$; $\psi_{i+1} \leftarrow \psi^{\text{cur}}$; $\tau_{i+1} \leftarrow \tau^{\text{cur}}$;
end

Algorithm 18: Random scan Gibbs sampler.

Input: An initial value $\theta^{(0)}$ such that $f_{\Theta|Y}(\theta^{(0)} \mid y) > 0$, the number of iterations N and a probability vector $\beta_1, \ldots, \beta_d$: each $\beta_j > 0$ and $\beta_1 + \cdots + \beta_d = 1$.

Result: Values simulated from a Markov chain which has the posterior distribution as its invariant distribution.

$\theta^{\text{cur}} \leftarrow \theta^{(0)}$;
for $i = 0, 1, \ldots, N$ **do**
  select j from $\{1, \ldots, d\}$ with probabilities $(\beta_1, \ldots, \beta_d)$;
  draw a new value for the component θ_j^{cur} from the posterior full conditional $f_{\Theta_j|\Theta_{-j},Y}(\theta_j \mid \theta_{-j}^{\text{cur}}, y)$;
  $\theta^{(i+1)} \leftarrow \theta^{\text{cur}}$;
end

need to choose the proposal distribution as in the Metropolis–Hastings sampler; the proposals of the Gibbs sampler are somehow automatically tuned to the target posterior. However, if some of the components of θ are strongly correlated in the posterior, then the convergence of the Gibbs sampler suffers. So one might want to reparametrize the model so that the transformed parameters are independent in their posterior. Unfortunately, most reparametrizations destroy the conditional conjugacy properties on which the attractiveness of the Gibbs sampler depends.

The name Gibbs sampling is actually not quite appropriate. Gibbs studied distributions arising in statistical physics (often called Gibbs distributions or Boltzmann distributions), which have densities of the form

$$f(x_1, \ldots, x_d) \propto \exp\left(-\frac{1}{kT} E(x_1, \ldots, x_d)\right),$$

where $(x_1, \ldots, x_d)$ is the state of physical system, k is a constant, T is the temperature of the system, and $E(x_1, \ldots, x_d) > 0$ is the energy of the system. The Geman brothers used a computational method (simulated annealing), where a computational parameter corresponding to the the temperature of a Gibbs distribution was gradually lowered towards zero. At each temperature the distribution of the system was simulated using the Gibbs sampler. This way they could obtain the configurations of minimal energy in the limit. The name Gibbs sampling was selected in order to emphasize the relationship with the Gibbs distributions. However, when the Gibbs sampler is applied to posterior inference, the temperature parameter is not needed, and therefore the reason for the name Gibbs has disappeared. Many authors have pointed this out this deficiency and proposed alternative names for the sampling method, but none of them have stuck.

7.6 Componentwise updates in the Metropolis–Hastings algorithm

Already Metropolis *et al.* and Hastings pointed out that one can use componentwise updates in the Metropolis–Hastings algorithm. This is sometimes called single-site updating or blockwise updating.

The parameter vector is divided into d components (or blocks)

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d),$$

which need not be scalars. In addtion, we need d proposal densities

$$\theta_j' \mapsto q_j(\theta_j' \mid \theta^{\text{cur}}), \qquad j = 1, \ldots, d,$$

which may all be different.

When it is time to update the jth component, we do a single Metropolis–Hastings step. When the current value of the parameter vector is θ^{cur}, we propose the vector θ', where the jth component is drawn from the proposal density $q_j(\theta_j \mid \theta^{\text{cur}})$, and the rest of the components of θ' are equal to those of the current value θ^{cur}. Then the proposal is accepted or rejected using the M–H ratio

$$r = \frac{p(\theta' \mid y)\, q_j(\theta_j^{\text{cur}} \mid \theta')}{p(\theta^{\text{cur}} \mid y)\, q_j(\theta_j' \mid \theta^{\text{cur}})} \tag{7.9}$$

The vectors θ' and θ^{cur} differ only in the jth place, and therefore one can write the M–H ratio (for updating the jth component) also in the form

$$r = \frac{p(\theta'_j \mid \theta^{\text{cur}}_{-j}, y)\, q_j(\theta^{\text{cur}}_j \mid \theta')}{p(\theta^{\text{cur}}_j \mid \theta^{\text{cur}}_{-j}, y)\, q_j(\theta'_j \mid \theta^{\text{cur}})}, \tag{7.10}$$

where we used the multiplication rule to express the joint posterior as

$$p(\theta \mid y) = p(\theta_{-j} \mid y)\, p(\theta_j \mid \theta_{-j}, y)$$

both in the numerator and in the denominator, and then cancelled the common factor $p(\theta^{\text{cur}}_{-j} \mid y)$. Although eqs. (7.9) and (7.10) are equivalent, notice that in eq. (7.9) we have the M–H ratio when we regard the joint posterior as the target distribution, but in eq. (7.10) we have ostensibly the M–H ratio, when the target is the posterior full conditional of component j. If one then selects as q_j the posterior full conditional of the component θ_j for each j, then each proposal is accepted and the Gibbs sampler ensues.

One can use this procedure either a systematic or a random scan sampler, as is the case with the Gibbs sampler. The resulting algorithm is often called the Metropolis–within–Gibbs sampler. (The name is illogical: the Gibbs sampler is a special case of the Metropolis–Hastings algorithm with componentwise updates.) This is also a very popular MCMC algorithm, since then one does not have to design a single complicated multivariate proposal density but p simpler proposal densities, many of which may be full conditional densities of the posterior.

Small modifications in the implementation can sometimes make a big difference to the efficiency of the sampler. One important decision is how to divide the parameter vector into components. This is called **blocking** or **grouping**. As a general rule, the less dependent the different components are in the posterior, the better the sampler. Therefore it may be a good idea to combine highly correlated components into a single block, with is then updated as a single entity.

It is sometimes useful to update the whole vector jointly using a single Metropolis–Hastings acceptance test, even if the proposed value is build up component by component taking advantage of conditional conjugacy properties. These and other ways of improving the performance of MCMC algorithms in the context of specific statistical models are topics of current research.

7.7 Analyzing MCMC output

After the MCMC algorithm has been programmed and tested, the user should investigate the properties of the algorithm for the particular problem he or she is trying to solve. There are available several tools, e.g., for

- diagnosing convergence
- estimating Monte Carlo standard errors.

We discuss some of the simpler tools.

A **trace plot** of a parameter ϕ is a plot of the iterates $\phi^{(t)}$ against the iteration number t. These are often examined for each of the components of the

parameter vector, and sometimes also for selected scalar functions of the parameter vector. A trace plot is also called a *sample path*, a *history plot* or a *time series plot*. If the chain mixes well, then the trace plots move quickly away from their starting values and they wiggle vigorously in the region supported by the posterior. In that case one may select the length of the burn-in by examining trace plots. (This is not foolproof, since the chain may only have converged momentarily to some neighborhood of a local maximum of the posterior.) If the chain mixes poorly, then the traces will remain nearly constant for many iterations and the state may seem to wander systematically towards some direction. Then one may need a huge number of iterations before the traces show convergence.

An **autocorrelation plot** is a plot of the autocorrelation of the sequence $\phi^{(t)}$ at different iteration lags, where ϕ may denote any of the components of the parameter vector, or any interesting function of the parameter vector. The autocorrelation function (acf) of a stationary sequence of RVs (X_i) at lag k is defined by

$$R(k) = \frac{E[(X_i - \mu)(X_{i+k} - \mu)]}{\sigma^2}, k = 0, 1, 2, \ldots,$$

where $\mu = EX_i$, $\sigma^2 = \operatorname{var} X_i$, and the assumption of stationarity entails that μ, σ^2 and $R(k)$ do not depend on index i. For an i.i.d. sequence the autocorrelation function is one at lag zero and zero otherwise. When dealing with MCMC output, one needs to estimate the autocorrelation function, and then to plot the estimate. These autocorrelation plots can be produced for all the interesting components of θ, but one should reject the burn-in before estimating the autocorrelation so that one analyzes only that part of the history where the chain is approximately stationary. A chain that mixes slowly exhibits slow decay of the autocorrelation as the lag increases. When there are more than one parameter, one may also examine cross-correlations between the parameters.

There exist tools for **convergence diagnostics**, which try to help in deciding whether the chain has already approximately reached its stationary distribution and in selecting the length of the burn-in period. E.g., in the approach of Gelman and Rubin, the chain is run many times starting from separate starting values dispersed over the support of the posterior. After the burn-in has been discarded, one calculates statistics which try to check whether all the chains have converged to the same distribution. In some other approches one needs to simulate only a single chain and one compares the behaviour of the chain in the beginning and in the end of the simulation run. Such convergence diagnostic are available in the `coda` R package and in the `boa` R package. However, convergence diagnostic tools can not prove that the chain has converged. They only help you to detect obvious cases of non-convergence.

If the chain seems to have converged, then it is of interest to estimate standard errors for the scalar parameters. The naive estimate (which is correct for i.i.d. sampling) would be to calculate the sample standard deviation of the last L iterations divided by $\sqrt{L}$ (after the burn-in has been discarded). However, MCMC iterates are typically positively correlated, and therefore this would underestimate the standard error severely.

A simple method for estimating the standard errors for posterior expectations

$$E[h(\Theta) \mid Y = y]$$

is the method of **batch means** [9], where the L last iterates are divided into a non-overlapping batches of length b. Then one computes the mean $\bar{h}_j$ of the values $h(\theta^{(t)})$ inside each of the batches $j = 1, \ldots, a$ and estimates the standard error of the grand mean $\bar{h}$ as the square roof of

$$\frac{1}{a} \frac{1}{a-1} \sum_{j=1}^{a} (\bar{h}_j - \bar{h})^2,$$

where $\bar{h}$ is the grand mean calculated from all the the L last iterates $h(\theta^{(t)})$. The idea here is to treat the batch means as i.i.d. random variables whose expected value is the posterior expectation. One should perhaps select the batch length as a function of the simulation length, e.g., with the rule $b = \lfloor \sqrt{L} \rfloor$.

There are, however, more sophisticated methods available for estimating the standard errors in MCMC. Some of these are available in the `boa` and `coda` R packages.

7.8 Example

Consider the two dimensional normal distribution $N(0, \Sigma)$ as the target distribution, where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad -1 < \rho < 1, \quad \sigma_1, \sigma_2 > 0,$$

and ρ is nearly one. Of course, it is possible to sample this two-variate normal distribution directly. However, we next apply MCMC algorithms to this highly correlated toy problem in order to demonstrate properties of the Gibbs sampler and a certain Metropolis–Hastings sampler.

The full conditionals of the target distribution are given by

$$\begin{aligned} [\Theta_1 \mid \Theta_2 = \theta_2] &\sim N\left(\tfrac{\rho\sigma_1}{\sigma_2}\theta_2, (1-\rho^2)\sigma_1^2\right) \\ [\Theta_2 \mid \Theta_1 = \theta_1] &\sim N\left(\tfrac{\rho\sigma_2}{\sigma_1}\theta_1, (1-\rho^2)\sigma_2^2\right), \end{aligned}$$

and these are easy to simulate. We now suppose that

$$\rho = 0.99, \quad \sigma_1 = \sigma_2 = 1.$$

Figure 7.2 shows the ten first steps of the Gibbs sampler, when all the component updates ("half-steps" of the sampler) are shown. Since ρ is almost one, the Gibbs sampler is forced to take small steps, and it takes a long time for it to explore the main support of the target distribution.

Another strategy would be to generate the proposal in two stages as follows. We first draw θ_1' from some convenient proposal distribution, e.g., by the random walk proposal

$$\theta_1' = \theta_1^{\text{cur}} + w,$$

where w is generated from (say) $N(0, 4)$. Then we draw θ_2' from the full conditional distribution of θ_2 conditioning on the proposed value θ_1'. Then the overall proposal density is given by

$$q((\theta_1', \theta_2') \mid (\theta_1^{\text{cur}}, \theta_2^{\text{cur}})) = N(\theta_1' - \theta_1^{\text{cur}} \mid 0, 4)\, N(\theta_2' \mid \frac{\rho\sigma_2}{\sigma_1}\theta_1', (1-\rho^2)\sigma_2^2)$$

Figure 7.2: The first ten iterations of the Gibbs sampler. The three contour lines enclose 50 %, 90 % and 99 % of the probability mass of the target distribution.

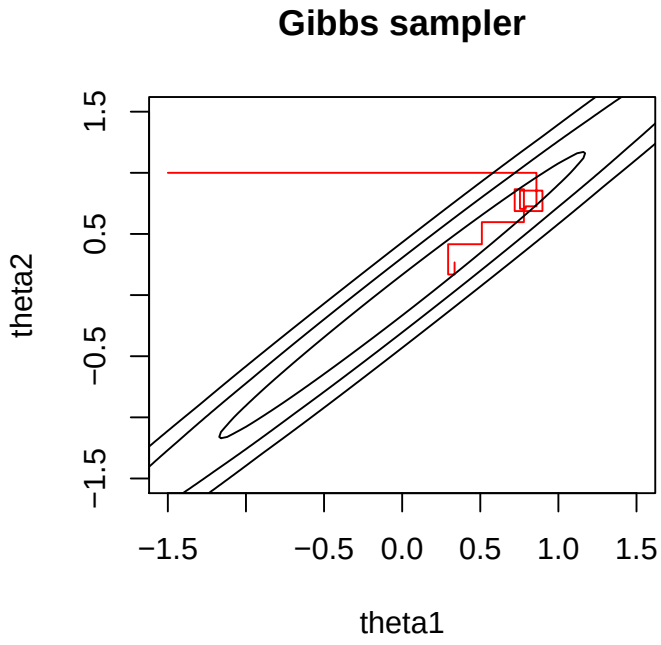


Figure 7.3: The first ten iterations of the Metropolis–Hastings sampler. Notice the sampler produced less than ten distinct θ values. The three contour lines enclose 50 %, 90 % and 99 % of the probability mass of the target distribution.

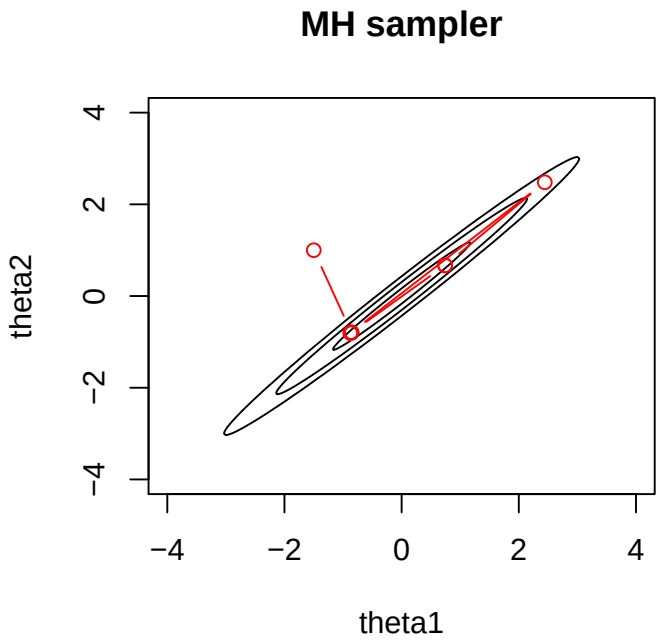
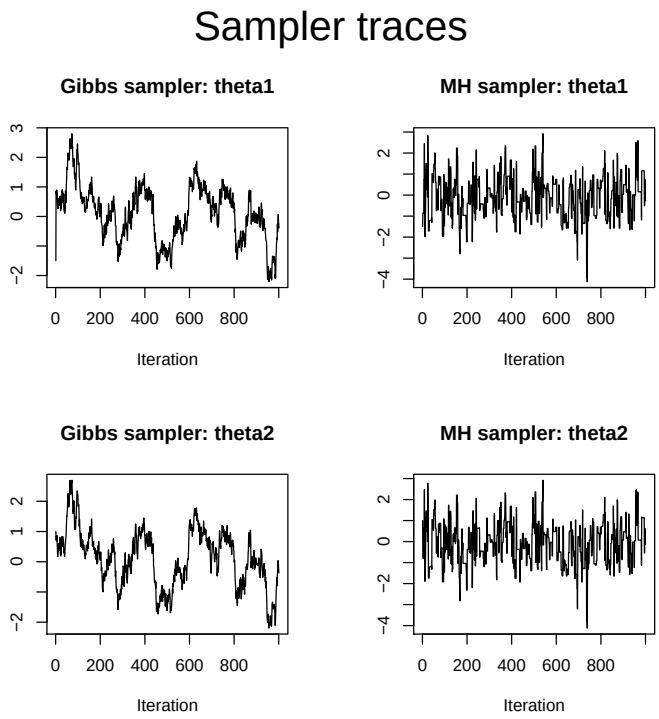


Figure 7.4: Sampler traces for the two components θ_1 and θ_2 using the Gibbs sampler and the Metropolis–Hastings sampler.



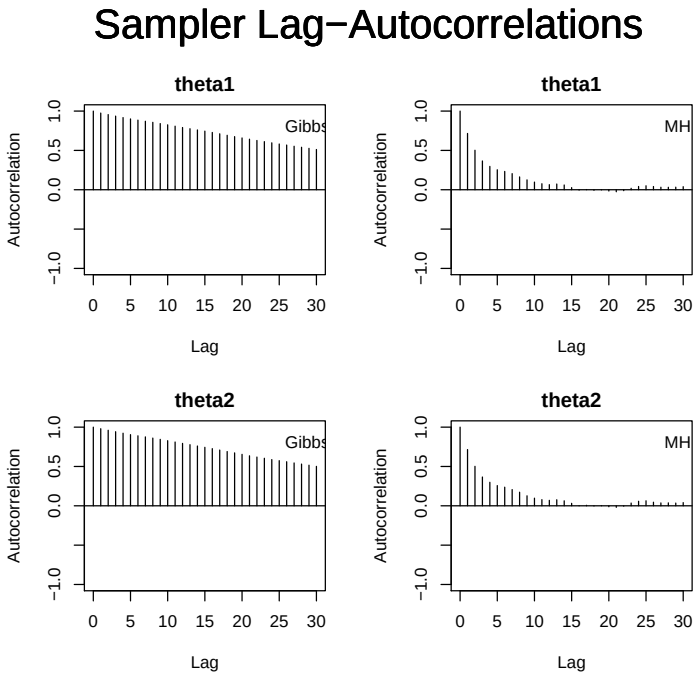
We then either accept or reject the transition from θ^{cur} to θ' using the ordinary acceptance rule of the Metropolis–Hastings sampler. This algorithm explores the target distribution much more efficiently, as can be guessed from Figure 7.3, which shows the first ten iterations of the sampler. The random walk proposal gives the component θ_1 freedom to explore the parameter space, and then the proposal from the full conditional for θ_2 draws the proposed pair into the main support of the target density.

Figure 7.4 shows the traces of the components using the two algorithms. The Metropolis–Hastings sampler seems to mix better than the Gibbs sampler, since there seems to be less dependence between the consecutive simulated values. Figure 7.5 shows the autocorrelation plots for the two components using the two different samplers. The autocorrelation functions produced by the Gibbs sampler decay more slowly than those produced by the Metropolis–Hastings sampler, and this demonstrates that we obtain better mixing with the Metropolis–Hastings sampler.

7.9 Literature

The original references on the Metropolis sampler, the Metropolis–Hastings sampler and the Gibbs sampler are [10, 8, 5]. The article by Gelfand and Smith [4] finally convinced the statistical community about the usefulness of these meth-

Figure 7.5: Sampler autocorrelation functions for the two components θ_1 and θ_2 using the Gibbs sampler and the Metropolis–Hastings sampler.



ods in Bayesian inference. The books [6, 3] contain lots of information on MCMC methods and their applications.

The books by Nummelin [12] or Meyn and Tweedie [11] can be consulted for the theory of Markov chains in a general state space. The main features of the general state space theory are explained in several sources, including [2, Ch. 14] or [13, Ch. 6].

Bibliography

[1] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373, 2008.

[2] Krishna B. Athreya and Soumendra N. Lahiri. *Measure Theory and Probability Theory*. Springer Texts in Statistics. Springer, 2006.

[3] Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[4] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[7] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

[9] Averll M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd edition, 1991.

[10] N. Metropolis, A. Rosenbluth, , M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[11] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.

[12] Esa Nummelin. *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge University Press, 1984.

[13] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

[14] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[15] Jeffrey S. Rosenthal. Optimal proposal distributions and adaptive MCMC. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC Press, 2011.

Chapter 8

Auxiliary Variable Models

8.1 Introduction

We are interested in an actual statistical model, with joint distribution

$$\mathrm{p}_{\mathrm{act}}(y, \theta) = \mathrm{p}_{\mathrm{act}}(y \mid \theta)\ \mathrm{p}_{\mathrm{act}}(\theta),$$

but where the posterior $\mathrm{p}_{\mathrm{act}}(\theta \mid y)$ is awkward to sample from. Suppose we are able to reformulate the original model by introducing a new random variable Z such that the marginal distribution of (y, θ) in the new model is the same as the joint distribution of (y, θ) in the original model, i.e., we assume that

$$\int \mathrm{p}_{\mathrm{aug}}(y, \theta, z)\, \mathrm{d}z = \mathrm{p}_{\mathrm{act}}(y, \theta). \tag{8.1}$$

When this is the case, we can forget the distinction between the actual model $\mathrm{p}_{\mathrm{act}}(\cdot)$ and the augmented model $\mathrm{p}_{\mathrm{aug}}(\cdot)$ and use the generic symbol $p(\cdot)$ to denote the densities calculated under either of the models. Here the *augmentation parameter*, the *auxiliary variable*, the *latent variable* or the *latent data* Z can be anything. However, it requires ingenuity and insight to come up with useful auxiliary variables.

Sometimes it is possible to sample much more efficiently from $p(\theta, z \mid y)$ than from $p(\theta \mid y)$. In such a case we can sample from the posterior $p(\theta, z \mid y)$, and we get a sample from the marginal posterior of θ by ignoring the z components of the (θ, z) sample. If both the full conditionals $p(\theta \mid z, y)$ and $p(z \mid \theta, y)$ are available in the sense that we know how to sample from these distribution, then implementing the Gibbs sampler is straightforward.

8.2 Slice sampler

Suppose we want to simulate from a distribution having the unnormalized density $q(\theta)$. By the fundamental theorem of simulation, this is equivalent to simulating (θ, z) from the uniform distribution under the graph of q, i.e., from $\mathrm{Uni}(A)$, the uniform distribution on the set

$$A = \{(\theta, z) : 0 < z < q(\theta)\}.$$

This distribution has the unnormalized density

$$p(\theta, z) \propto 1_A(\theta, z) = 1_{(0,q(\theta))}(z) = 1(0 < z < q(\theta))$$

The full conditional of Z is proportional to the joint density, considered as a function of z, i.e.,

$$p(z \mid \theta) \propto p(\theta, z) \propto 1(0 < z < q(\theta)),$$

and this an unnormalized density of the uniform distribution on the interval $(0, q(\theta))$.

Similarly, the full conditional of θ is the uniform distribution on the set (depending on z), where

$$1(0 < z < q(\theta)) = 1,$$

since the joint density is constant on this set. That is, the full conditional of θ is the uniform distribution on the set

$$B(z) = \{\theta : q(\theta) > z\}.$$

The resulting Gibbs sampler is called the slice sampler (for the distribution determined by q). The slice sampler is attractive, if the uniform distribution on the set $B(z)$ is easy to simulate.

Example 8.1. Let us consider the truncated standard normal distribution corresponding to the unnormalized density

$$q(\theta) = \exp\left(-\frac{1}{2}\theta^2\right) 1_{(\alpha,\infty)}(\theta),$$

where the truncation point $\alpha > 0$.

We can get a correlated sample $\theta_1, \theta_2, \ldots$ from this distribution as follows.

1. Pick an initial value $\theta_0 > \alpha$.
2. For $i = 1, 2 \ldots$
   - Draw z_i from $\mathrm{Uni}(0, q(\theta_{i-1}))$.
   - Draw θ_i from $\mathrm{Uni}(\alpha, \sqrt{-2 \ln z_i})$.

△

Simulating the uniform in the set $B(z)$ may turn out to be unwieldy. Usually, the target density can be decomposed into a product of functions,

$$p(\theta \mid y) \propto \prod_{i=1}^{n} q_i(\theta).$$

Then one may try the associated augmentation, where one introduces n auxiliary variables Z_i such that, conditionally on θ, the Z_i have independently the uniform distribution on $(0, q_i(\theta))$. In the augmented model, the full conditional of θ is the uniform distribution on the set

$$C(z) = \cap_{i=1}^{n} \{\theta : q_i(\theta) > z_i\},$$

and this may be easier to simulate. Typically, the more auxiliary variables one introduces, the slower is the mixing of the resulting chain.

8.3 Missing data problems

In many experiments the posterior distribution is easy to summarize if all the planned data are available. However, if some of the observations are missing, then the posterior is more complex. Let Z be the missing data and let y be the observed data. The full conditional

$$p(\theta \mid z, y)$$

is the posterior from the complete data, and it is of a simple form (by assumption). Often also the full conditional of the missing data

$$p(z \mid \theta, y)$$

is easy to sample from. Then it is straightforward to use the Gibbs sampler.

Here the joint distribution in the reformulated model is

$$\mathrm{p}_{\mathrm{aug}}(y, \theta, z) = \mathrm{p}_{\mathrm{act}}(\theta)\ \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta).$$

In order to check the equivalence of the original and of the reformulated model, see (8.1), it is sufficient to check that

$$\int \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta)\, \mathrm{d}z = \mathrm{p}_{\mathrm{act}}(y \mid \theta).$$

This is trivial, if the complete data likelihood is specified in the form

$$\mathrm{p}_{\mathrm{aug}}(y, z \mid \theta) = \mathrm{p}_{\mathrm{act}}(y \mid \theta)\ \mathrm{p}_{\mathrm{aug}}(z \mid y, \theta).$$

However, often the complete data likelihood is specified in some other way, and then checking the equivalence to the original model requires more thought.

Example 8.2. Let us consider the famous genetic linkage example, where we have the multinomial likelihood

$$p(y \mid \theta) = \mathrm{Mult}\left((y_1, y_2, y_3, y_4) \mid n, \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)\right).$$

Here $0 < \theta < 1$, and $y = (y_1, y_2, y_3, y_4)$, where the y_j:s are the observed frequencies of the four categories. We take the uniform prior $\mathrm{Uni}(0, 1)$ for θ. The posterior is proportional to

$$q(\theta) = \theta^{y_4}\,(1-\theta)^{y_2+y_3}\,(2+\theta)^{y_1}, \qquad 0 < \theta < 1$$

but thanks to the last factor, this is not of a standard form.

However, suppose that the first category with frequency y_1 is an amalgamation of two subclasses with probabilities $\theta/4$ and $1/2$, but the distinction between the subclasses has not been observed. Let Z be the frequency of the first subclass (with class probability $\theta/4$). Then the frequency of the second subclass (with class probability $1/2$) is $y_1 - Z$. Our reformulated model states that

$$p(z, y \mid \theta) = p(z, y_1, y_2, y_3, y_4 \mid \theta) = \\ \mathrm{Mult}\left((z, y_1 - z, y_2, y_3, y_4) \mid n, \left(\frac{1}{4}\theta, \frac{1}{2}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right)\right)$$

Let us check that the reformulated model and the original model are equivalent. If we combine the frequencies X_{11} and X_{12} in the the multinomial distribution

$$(X_{11}, X_{12}, X_2, X_3, X_4) \sim \text{Mult}(n, (p_{11}, p_{12}, p_2, p_3, p_4)),$$

then we obtain the multinomial distribution

$$(X_{11} + X_{12}, X_2, X_3, X_4) \sim \text{Mult}(n, (p_{11} + p_{12}, p_2, p_3, p_4)),$$

and this is obvious when one thinks of the repeated sampling definition of the multinomial distribution. This shows that our original model and the reformulated model are equivalent.

The posterior of θ given the complete data consisting of y and z is given by

$$\begin{aligned} p(\theta \mid y, z) &\propto p(y, z \mid \theta)\, p(\theta) \\ &\propto \left(\frac{1}{4}\theta\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}\theta\right)^{y_4} \\ &\propto \theta^{z+y_4}\,(1-\theta)^{y_2+y_3}. \end{aligned}$$

This is an unnormalized density of the beta distribution $\text{Be}(z+y_4+1, y_2+y_3+1)$, which can be sampled directly.

The full conditional of Z is trickier to recognize. Notice that Z is an integer such that $0 \le Z \le y_1$. It is critical to notice that the normalizing constant of the multinomial pmf $p(z, y \mid \theta)$ depends on z. While you can omit from the likelihood any terms which depend only on the *observed* data, you must keep those terms which depend on the unknowns: parameters or *missing* data.

As a function of z,

$$\begin{aligned} p(z \mid \theta, y) &\propto p(z, y \mid \theta)\, p(\theta) = p(z, y \mid \theta) \\ &= \frac{n!}{z!\,(y_1 - z)!\,y_2!\,y_3!\,y_4!} \left(\frac{1}{4}\theta\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}\theta\right)^{y_4} \\ &\propto \frac{y_1!}{z!\,(y_1 - z)!} \left(\frac{\theta}{4}\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \\ &= \binom{y_1}{z} \left(\frac{\frac{\theta}{4}}{\frac{\theta}{4} + \frac{1}{2}}\right)^z \left(\frac{\frac{1}{2}}{\frac{\theta}{4} + \frac{1}{2}}\right)^{y_1 - z} \left(\frac{\theta}{4} + \frac{1}{2}\right)^{z + y_1 - z} \\ &\propto \binom{y_1}{z} \left(\frac{\theta}{2+\theta}\right)^z \left(1 - \frac{\theta}{2+\theta}\right)^{y_1 - z}, \qquad z = 0, 1, \ldots, y_1. \end{aligned}$$

From this we see that the full conditional of Z is the binomial $\text{Bin}(y_1, \theta/(2+\theta))$, which we also are able to simulate directly. Gibbs sampling in the reformulated model is straightforward. △

8.4 Probit regression

We now consider a regression model, where each of the responses is binary: zero of one. In other words, each of the responses has the Bernoulli distribution (the binomial distribution with sample size one). Conditionally on the parameter

vector θ, the responses Y_i are assumed to be independent, and Y_i is assumed to have success probability

$$q_i(\theta) = P(Y_i = 1 \mid \theta),$$

which is a function of the parameter vector θ. That is, the model assumes that

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(q_i(\theta)), \qquad i = 1, \ldots, n,$$

where $B(p)$ is the Bernoulli distribution with success probability $0 \leq p \leq 1$.

We assume that the success probability of the i'th response depends on θ and on the value of the covariate vector x_i for the i'th case. The covariate vector consists of observed characteristics which might influence the probability of success. We would like to model the success probability in terms of a linear predictor, which is the inner product $x_i^T\theta$ of the covariate vector and the parameter vector. For instance, if we have observed a single explanatory scalar variable t_i connected with the response y_i, then the linear predictor could be

$$x_i^T\theta = \alpha + \beta\, t_i, \qquad x_i = (1, t_i), \quad \theta = (\alpha, \beta).$$

Notice that we typically include the constant "1" in the covariate vector.

The linear predictor is not constrained to the range $[0, 1]$ of the probability parameter, and therefore we need to map the values of the linear predictor into that range. The standard solution is to posit that

$$q_i(\theta) = F(x_i^T\theta), \qquad i = 1, \ldots, n.$$

where F is the cumulative distribution function of some continuous distribution supported on the whole real line. Since $0 \leq F \leq 1$, here $q_i(\theta)$ is a valid probability parameter for the Bernoulli distribution for any value of θ. Such a binary regression model belongs to the class of generalized linear models (GLMs). In this context, the inverse function of the cdf F is called the *link function*; the model is linear on the scale of the link function,

$$F^{-1}(q_i(\theta)) = x_i^T\theta, \qquad i = 1, \ldots, n.$$

The most popular popular choice for the link function in binary regression is the logit link $\ln(q/(1-q))$, which corresponds to choosing F to be the cdf of the logistic distribution,

$$F(u) = \frac{\mathrm{e}^u}{1 + \mathrm{e}^u} = \text{logit}^{-1}(u).$$

The logit link has a special status in binary regression, since the logit link happens to be what is known as the canonical link function in the theory of generalized linear models. In *probit regression* we take $F = \Phi$, where Φ is the cdf of the standard normal $N(0, 1)$, i.e., we assume that

$$q_i(\theta) = P(Y_i = 1 \mid \theta) = \Phi(x_i^T\theta), \qquad i = 1, \ldots, n.$$

The third commonly used link function is the complementary log-log link

$$F^{-1}(q) = \ln(-\ln(1-q)), \qquad 0 < q < 1,$$

which corresponds to the cdf $F(u) = 1 - \exp(-\exp(u))$. The maximum likelihood estimate (MLE) for binary regression using either logit, probit or complementary log-log link can be calculated with standard software, e.g., using the function `glm` of R.

We can write the likelihood in binary regression immediately, i.e.,

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} F(x_i^T\theta)^{y_i} \left(1 - F(x_i^T\theta)\right)^{1-y_i}. \tag{8.2}$$

Posterior inference can be based directly on this expression. Gibbs sampling seems impossible, but a suitable MCMC algorithm could be, e.g., the independence sampler with a multivariate Student's t distribution, whose center and covariance matrix are selected based on the MLE and its approximate covariance matrix, which can be calculated with standard software.

From now on, we will discuss the probit regression model, and its well-known auxiliary variable reformulation, due to Albert and Chib [1]. The likelihood is given in equation (8.2) with F equal to Φ, and our prior is the normal distribution with mean μ_0 and precision matrix Q_0,

$$p(\theta) = N(\theta \mid \mu_0, Q_o^{-1}).$$

Let us introduce n latent variables (i.e., unobserved random variables)

$$[Z_i \mid \theta] \overset{\text{ind}}{\sim} N(x_i^T\theta, 1), \qquad i = 1, \ldots, n.$$

This notation signifies that the Z_i's are independent, conditionally on θ. We may represent the latent variables Z_i using n i.i.d. random variables $\epsilon_i \sim N(0,1)$ (which are independent of everything else),

$$Z_i = x_i^T\theta + \epsilon_i, \qquad i = 1, \ldots, n.$$

Consider n RVs Y_i which are defined by

$$Y_i = 1(Z_i > 0) = \begin{cases} 1, & \text{when } Z_i > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Conditionally on θ, the random variables Y_i are independent, Y_i takes on the value zero or one, and

$$P(Y_i = 1 \mid \theta) = P(Z_i > 0 \mid \theta) = P(x_i^T\theta + \epsilon_i > 0) = P(-\epsilon_i < x_i^T\theta) = \Phi(x_i^T\theta).$$

Here we used the fact that $-\epsilon_i \sim N(0,1)$ which follows from the symmetry of the standard normal. Therefore the marginal distribution of $Y = (Y_1, \ldots, Y_n)$ given θ is the same as in the original probit regression model. Our reformulated model has the structure

$$\mathrm{p}_{\mathrm{aug}}(y, \theta, z) = \mathrm{p}_{\mathrm{act}}(\theta)\ \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta),$$

and we have just argued that

$$\int \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta)\, \mathrm{d}z = \mathrm{p}_{\mathrm{act}}(y \mid \theta).$$

This shows that our reformulated model is equivalent with the original probit regression model.

The reformulated probit regression model has the following hierarchical structure,

$$\begin{aligned} \Theta &\sim N(\mu_0, Q_0^{-1}) && (8.3)\\ [Z \mid \Theta = \theta] &\sim N(X\theta, I) && (8.4)\\ Y &= 1_+(Z), && (8.5) \end{aligned}$$

where X is the known design matrix with ith row equal to x_i^T, Z is the column vector of latent variables, and $1_+(Z)$ means the vector

$$1_+(Z) = \begin{bmatrix} 1(Z_1 > 0) \\ \vdots \\ 1(Z_n > 0) \end{bmatrix},$$

where we write $1(Z_i > 0)$ for the indicator $1_{(0,\infty)}(Z_i)$. Therefore we can regard the original probit regression model as a missing data problem where we have a normal regression model on the latent data $Z = (Z_1, \ldots, Z_n)$ and the observed responses Y_i are incomplete in that we only observe whether $Z_i > 0$ or $Z_i \leq 0$.

The joint distribution of the reformulated model can be expressed as

$$p(\theta, y, z) = p(\theta)\, p(z \mid \theta)\, p(y \mid z),$$

where

$$p(y \mid z) = \prod_{i=1}^{n} p(y_i \mid z_i),$$

and further

$$p(y_i \mid z_i) = 1(z_i > 0)\, 1(y_i = 1) + 1(z_i \leq 0)\, 1(y_i = 0).$$

(Y_i is a deterministic function of Z_i. The preceding representation is possible, since Y_i has a discrete distribution.)

The full conditional of θ is easy to calculate, since

$$p(\theta \mid z, y) \propto p(\theta, y, z) \propto p(\theta)\, p(z \mid \theta),$$

but this is the same as the posterior in a certain linear regression model, namely the multivariate normal $N(\theta \mid \mu_1, Q_1^{-1})$, whose parameters μ_1 and Q_1 can be solved from

$$Q_1 = Q_0 + X^T X, \qquad Q_1\, \mu_1 = Q_0\, \mu_0 + X^T z.$$

The other full conditional distribution is also easy to derive. As a function of z, we have

$$p(z \mid \theta, y) \propto p(z \mid \theta)\, p(y \mid z) = \prod_{i=1}^{n} N(z_i \mid x_i^T \theta, 1)\, p(y_i \mid z_i)$$

This is a distribution, where the components Z_i are independent, and follow truncated normal distributions, i.e.,

$$\begin{aligned} [Z_i \mid \theta, y] &\sim N(x_i^T \theta, 1)\, 1(Z_i > 0), && \text{if } y_i = 1,\\ [Z_i \mid \theta, y] &\sim N(x_i^T \theta, 1)\, 1(Z_i \leq 0), && \text{if } y_i = 0. \end{aligned}$$

Notice that the side of the truncation for Z_i depends on the value of the binary response y_i. Simulating the full conditional distribution $p(z \mid \theta, y)$ is also straightforward, since we only have to draw independently n values from truncated normal distributions with known parameters and known semi-infinite truncation intervals. Since all the needed full conditional distributions are easily simulated, implementing the Gibbs sampler is straightforward in the latent variable reformulation.

What is the practical benefit of the latent variable reformulation of the probit regression model? In the original formulation of the probit regression model, the components of θ are dependent in their posterior. MCMC sampling will be inefficient unless we manage to find a proposal distribution which is adapted to the form of the posterior distribution. After the reformulation, Gibbs sampling becomes straightforward. In the latent variable reformulation, most of the dependencies in the posterior are transferred to the multivariate normal distribution $p(\theta \mid z, y)$, where they are easy to handle. The components of Z are independent in the other needed full conditional distribution $p(z \mid \theta, y)$.

8.5 Scale mixtures of normals

Student's t distribution with $\nu > 0$ degrees of freedom can be expressed as a scale mixture of normal distributions. Namely, if

$$\Lambda \sim \mathrm{Gam}(\nu/2, \nu/2), \quad \text{and} \quad [W \mid \Lambda = \lambda] \sim N(0, \frac{1}{\lambda}),$$

then the marginal distribution of W is t_ν. We can use this property to eliminate Student's t distribution (and the Cauchy distribution which is equal to t_1) from any statistical model.

Albert and Chib considered approximating the logit link with the t_ν link in binary regression. The logit link is already well approximated by the probit link in the sense that

$$\mathrm{logit}^{-1}(u) \approx \Phi\left(\sqrt{\frac{\pi}{8}}u\right),$$

when u is near zero. Here the scaling factor $\sqrt{\pi/8}$ has been selected so that the derivatives of the two curves are equal for $u = 0$. The approximation is not perfect away from zero. However, if one uses the distribution function F_ν of the t_ν distribution (e.g., with $\nu = 8$ degrees of freedom), then one can choose the value of the scaling factor s so that we have a much better approximation

$$\mathrm{logit}^{-1}(u) \approx F_\nu(su)$$

for all real u. Making use of the scaling factor s, we can switch between a logit regression model and its t_ν regression approximation.

We now consider, how we can reformulate the binary regression model which has the t_ν link, i.e.,

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(F_\nu(x_i^T\theta)), \qquad i = 1, \ldots, n. \tag{8.6}$$

Here the degrees of freedom parameter ν is fixed. Also this reformulation is due to Albert and Chib [1].

The first step is to notice that we can represent the responses as

$$Y_i = 1(Z_i > 0), \quad \text{where} \quad Z_i = x_i^T\theta + W_i, \qquad i = 1, \ldots, n,$$

where $W_i \sim t_\nu$ are i.i.d. and independent of everything else. This holds since

$$P(Z_i > 0 \mid \theta) = P(x_i^T\theta + W_i > 0) = P(-W_i < x_i^T\theta) = F_\nu(x_i^T\theta).$$

Here we used the fact that $-W_i \sim t_\nu$ which follows from symmetry of the t distribution. Besides, the Z_i's are independent, conditionally on θ. Next we eliminate the t_ν distribution by introducing n i.i.d. latent variables Λ_i, each having the $\text{Gam}(\nu/2, \nu/2)$ distribution. If we choose $N(\mu_0, Q_0^{-1})$ as the prior for Θ, then we end up with the following hierarchical model

$$\begin{aligned}
\Theta &\sim N(\mu_0, Q_0^{-1}), && (8.7)\\
\Lambda_i &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\nu/2, \nu/2), \qquad i = 1, \ldots, n && (8.8)\\
[Z \mid \Theta = \theta, \Lambda = \lambda] &\sim N\left(X\theta, [\text{diag}(\lambda_1, \ldots, \lambda_n)]^{-1}\right), && (8.9)\\
Y &= 1_+(Z). && (8.10)
\end{aligned}$$

This reformulation is equivalent with the original model (8.6).

The full conditionals in the reformulated model are easy to derive. The full conditional of θ is a multivariate normal. The full conditional of $\Lambda = (\Lambda_1, \ldots, \Lambda_n)$ is the distribution of n independent gamma distributed variables with certain parameters. The full conditional of Z is, once again, a distribution, where the components are independent and have truncated normal distributions.

Another well-known distribution, which can be expressed as a scale mixture of normal distributions is the Laplace distribution (the double exponential distribution), which has the density

$$\frac{1}{2}\mathrm{e}^{-|y|}, \qquad y \in \mathbb{R}\,.$$

If Y has the Laplace distribution, then it can be expressed as follows

$$V \sim \text{Exp}(1/2) \quad \text{and} \quad [Y \mid V = v] \sim N(0, v).$$

This relationship can be used to eliminate the Laplace distribution from any statistical model.

Even the logistic distribution with distribution function $\text{logit}^{-1}(z)$ can be expressed as a scale mixture of normals, but then one needs the Kolmogorov-Smirnov distribution, whose density and distribution function are, however, available only as series expansions. Using this device, one can reformulate the logistic regression model exactly using the Kolmogorov-Smirnov distribution, multivariate normal distribution and truncation, see Holmes and Held [3] for an implementation of the idea.

8.6 Literature

The slice sampler was proposed by Neal [4]. The data augmentation in the genetic linkage example is from the article by Tanner and Wong [6], who borrowed ideas from earlier work on the EM algorithm. The auxiliary variable

formulation of probit regression was proposed by Albert and Chib [1]. Also the reformulation of the t link is from this article. Scale mixtures of normals were characterized by Andrews and Mallows [2]. Tan, Tian and Ng [5] present many interesting computational approaches for Bayesian missing data problems.

Bibliography

[1] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36:99–102, 1974.

[3] Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006.

[4] Radford M. Neal. Slice sampling. *Annals of Statistics*, 23:705–767, 2003.

[5] Ming T. Tan, Guo-Liang Tian, and Kai Wang Ng. *Bayesian Missing Data Problems: EM, Data Augmentation and Noninterative Computation*. Chapman & Hall/CRC, 2010.

[6] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

Chapter 9

The EM Algorithm

The EM (Expectation–Maximization) algorithm is an iterative method for finding the mode of a marginal posterior density. It can also be used for finding the mode of a (marginalized) likelihood function. The idea is to replace the original maximization problem by a sequence of simpler optimization problems. In many examples the maximizers of the simple problems can be obtained in closed form.

Typically the EM algorithm is applied in an auxiliary variable (latent variable) formulation $p(y, \theta, z)$ of the original model $p(y, \theta)$, where θ is the parameter of interest, and Z is the auxiliary variable (or latent variable or missing data).

In the Bayesian framework, θ and z are unknown and y is observed, and the marginal posterior of θ in the joint posterior $p(\theta, z \mid y)$ is the posterior $p(\theta \mid y)$ of the original model, namely

$$p(\theta \mid y) = \int p(\theta, z \mid y) \, \mathrm{d}z.$$

In the frequentist framework, we consider z to be missing data and call $p(y, z \mid \theta)$ the complete-data likelihood. The observed-data likelihood $p(y \mid \theta)$ is found by marginalizing the complete-data likelihood

$$p(y \mid \theta) = \int p(y, z \mid \theta) \, \mathrm{d}z.$$

In an auxiliary variable formulation, the EM algorithm can be used to find either the posterior mode or the MLE (maximum likelihood estimate) of the original model.

9.1 Formulation of the EM algorithm

Let Z be the auxiliary variable and θ the parameter of interest. The EM algorithm can be formulated either for the mode of the marginal posterior of θ or for the mode of the observed-data likelihood of θ. In both cases one defines a function, usually called Q, which depends on two variables, θ and θ_0, where θ_0 stands for the current guess of the parameter vector θ_0. The function $Q(\theta \mid \theta_0)$ is defined as a certain expected value.

The EM algorithm alternates between two steps: first one calculates the Q function given the current guess θ_0 for the parameter vector (E-step), and then one maximizes $Q(\theta \mid \theta_0)$ with respect to θ in order to define the new guess for θ (M-step). This procedure is repeated until a fixed point of Q is obtained (or some other termination criterion is satisfied). This idea is formalized in algorithm 19. There $\arg\max$ denotes the maximizing argument (maximum point) of the function it operates on. If the maximizer is not unique, we may select any global maximizer.

Algorithm 19: The EM algorithm.

Input: An initial value $\theta^{(0)}$.

$k \leftarrow 0$;

repeat

(E-step) Calculate the function $Q(\theta \mid \theta^{(k)})$;

(M-step) Maximize $Q(\theta \mid \theta^{(k)})$ with respect to θ:

$$\theta^{(k+1)} \leftarrow \arg\max_{\theta} Q(\theta \mid \theta^{(k)})$$

Set $k \leftarrow k + 1$

until *the termination criterion is satisfied* ;

Return the last calculated value $\theta^{(k)}$;

Next we define the function Q for the two different objectives. When we want to calculate the mode of the (marginal) posterior density in a Bayesian model, we define $Q(\theta \mid \theta_0)$ as the expected value of the logarithm of the joint posterior density, conditioned on the data and on the current guess θ_0,

$$\begin{aligned} Q(\theta \mid \theta_0) &= E\left[\log p(\theta, Z \mid y) \mid \theta_0, y\right] \\ &= E\left[\log f_{\Theta,Z|Y}(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y\right] \\ &= \int \log f_{\Theta,Z|Y}(\theta, z \mid y)\, f_{Z|\Theta,Y}(z \mid \theta_0, y)\, \mathrm{d}z. \end{aligned} \tag{9.1}$$

The only random object in the above expected value is Z, and we use its distribution conditioned on the current value θ_0 and the data y.

When we want to calculate the mode of the observed-data likelihood in a frequentist model, we define $Q(\theta \mid \theta_0)$ as the expected complete-data log-likelihood, conditioning on the data and on the current value θ_0,

$$\begin{aligned} Q(\theta \mid \theta_0) &= E\left[\log p(y, Z \mid \theta) \mid \theta_0, y\right] \\ &= E\left[\log f_{Y,Z|\Theta}(y, Z \mid \theta) \mid \Theta = \theta_0, Y = y\right] \\ &= \int \log f_{Y,Z|\Theta}(y, z \mid \theta)\, f_{Z|\Theta,Y}(z \mid \theta_0, y)\, \mathrm{d}z. \end{aligned} \tag{9.2}$$

The Q function is defined as an expectation of a sum of a number terms. Luckily, we can treat all of the terms which do not depend on θ as constants. Namely, in the M-step we select a maximum point of the function $\theta \mapsto Q(\theta \mid \theta_0)$, and the ignored constants only shift the object function but do not change the location of the maximum point. That is, the functions

$$Q(\theta \mid \theta_0) \quad \text{and} \quad Q(\theta \mid \theta_0) + c(\theta_0, y)$$

achieve their maxima at the same points, when the "constant" $c(\theta_0, y)$ does not depend on the variable θ. In particular, we can ignore any factors which depend solely on the observed data y.

The maximization problem (M-step) can be solved in closed form in many cases where the joint posterior (or complete-data likelihood, respectively) belongs to the exponential family. Then the E- and M-steps boil down to the following steps: finding the expectations (given the current θ_0) of the sufficient statistics (which now depend on the missing data Z), and maximizing the resulting function with respect to the parameters θ.

If the maximizer cannot be solved analytically, then instead of the maximum point one can (in the M-step) select any value $\theta^{(k+1)}$ such that

$$Q(\theta^{(k+1)} \mid \theta^{(k)}) > Q(\theta^{(k)} \mid \theta^{(k)}).$$

The resulting algorithm is then called the generalized EM algorithm (GEM).

We will show later that the logarithm of the marginal posterior

$$\log f_{\Theta|Y}(\theta^{(k)} \mid y)$$

increases monotonically during the iterations of the EM or the GEM algorithms, if one defines Q by (9.1). On the other hand, if one defines Q by (9.2), then the observed-data log-likelihood

$$\log f_{Y|\Theta}(y \mid \theta^{(k)})$$

increases monotonically during the iterations. If these functions can be calculated easily, then a good check of the correctness of the implementation is to check that they indeed increase at each iteration.

Because of this monotonicity property, the EM algorithm converges to some local mode of the object function (except in some artificially constructed cases). If the object function has multiple modes, then one can try to find all of them by starting the EM iterations at many points scattered throughout the parameter space.

9.2 EM algorithm for probit regression

We return to the latent variable reformulation of the probit regression problem, i.e.,

$$\begin{aligned} \Theta &\sim N(\mu_0, R_0^{-1}) \\ [Z \mid \Theta = \theta] &\sim N(X\theta, I) \\ Y &= 1_+(Z), \end{aligned}$$

where X is the known design matrix, Z is the column vector of latent variables, and $1_+(Z)$ is the vector of indicators $1(Y_i > 0)$. We use the symbols ϕ and Φ for the density and df of the standard normal $N(0,1)$, and use R_0 to denote the precision matrix of the prior.

We have already obtained the distribution of the latent variables given θ and the data, $p(z \mid \theta, y)$. In it, the latent variables Z_i are independent and have the following truncated normal distributions

$$\begin{aligned} [Z_i \mid \theta, y] &\sim N(x_i^T\theta, 1)\, 1(Z_i > 0), \qquad \text{if } y_i = 1, \\ [Z_i \mid \theta, y] &\sim N(x_i^T\theta, 1)\, 1(Z_i \le 0), \qquad \text{if } y_i = 0. \end{aligned}$$

Now the joint posterior is

$$p(\theta, z \mid y) \propto p(y, \theta, z) = p(y \mid z)\, p(z \mid \theta)\, p(\theta).$$

Here $p(y \mid z)$ is simply the indicator of the constraints $y = 1_+(z)$. For any y and z values which satisfy the constraints $y = 1_+(z)$, the log joint posterior is given by

$$\begin{aligned}
\log p(\theta, z \mid y) &= \log p(z \mid \theta) + \log p(\theta) + c_1 \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2}(z - X\theta)^T (z - X\theta) + c_2 \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2} z^T z + \theta^T X^T z - \frac{1}{2}\theta^T X^T X \theta + c_2
\end{aligned}$$

where the constants c_i depends on the data y and the known hyperparameters, but not on z, θ or θ_0.

Since now

$$Q(\theta \mid \theta_0) = E[\log p(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y],$$

at first sight it may appear that we need to calculate both the expectations

$$v(\theta_0) = E[Z^T Z \mid \Theta = \theta_0, Y = y], \quad \text{and} \quad m(\theta_0) = E[Z \mid \Theta = \theta_0, Y = y],$$

but on further thought we notice that we actually need only the expectation $m(\theta_0)$. This is so, since the term containing $z^T z$ in $\log p(\theta, z \mid y)$ does not depend on θ. In the maximization of $Q(\theta \mid \theta_0)$ its expectation therefore only shifts the object function but does not affect the location of the maximizer.

Let us next solve the maximizer of $\theta \mapsto Q(\theta \mid \theta_0)$ and then check which quantities need to be calculated. In the following, c_i is any quantity, which does not depend on the variable θ (but may depend on y, θ_0 or the known hyperparameters).

$$\begin{aligned}
Q(\theta \mid \theta_0) &= E[\log p(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y] \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2}\theta^T X^T X \theta + \theta^T X^T m(\theta_0) + c_3 \qquad (9.3) \\
&= -\frac{1}{2}\theta^T (R_0 + X^T X)\theta + \theta^T \left[R_0 \mu_0 + X^T m(\theta_0) \right] + c_4
\end{aligned}$$

We now make the following observations.

1. The matrix $R_0 + X^T X$ is symmetric and positive definite. Symmetry is obvious, and for any $v \neq 0$,

   $$v^T (R_0 + X^T X) v = v^T R_0 v + v^T X^T X v > 0,$$

   since $v^T R_0 v > 0$ and $v^T X^T X v = (Xv)^T (Xv) \geq 0$.

2. If the matrix K is symmetric and positive definite, then the maximizer of the quadratic form

   $$-\frac{1}{2}(\theta - a)^T K (\theta - a)$$

   is a, since the quadratic form vanishes if and only if $\theta = a$.

3. The preceding quadratic form can developed as

$$-\frac{1}{2}(\theta - a)^T K(\theta - a) = -\frac{1}{2}\theta^T K\theta + \theta^T Ka + \text{constant}.$$

Therefore, the maximum point of

$$-\frac{1}{2}\theta^T K\theta + \theta^T b + c,$$

where K is assumed to be symmetric and positive definite, is

$$\theta = K^{-1}b.$$

(An alternative way to derive the formula for the maximum point is to equate the gradient $-K\theta + b$ of the quadratic function to the zero vector, and to observe that the Hessian $-K$ is negative definite.)

Based on the preceding observations, the maximizer of $\theta \mapsto Q(\theta \mid \theta_0)$ given in eq. (9.3) is given by

$$\theta_1 = (R_0 + X^T X)^{-1}(R_0\mu_0 + X^T m(\theta_0)). \tag{9.4}$$

However, we still need to calculate a concrete formula for the vector

$$m(\theta_0) = E[Z \mid \Theta = \theta_0, Y = y].$$

We need a formula for the expected value of the truncated normal distribution $N(\mu, \sigma^2)1_{(\alpha,\beta)}$ corresponding to the unnormalized density

$$f(v) \propto N(v \mid \mu, \sigma^2)1_{(\alpha,\beta)}(v) \tag{9.5}$$

where we can have $\alpha = -\infty$ or $\beta = \infty$. The moment generating function of this distribution is easy to calculate. Then we obtain its expected value (and higher moments, if need be) by differentiating the result.

Let Φ be the distribution function and ϕ the density function of the standard normal $N(0, 1)$. If V has the truncated normal distribution (9.5), then a simple calculation shows that

$$\begin{aligned} M(t) &= E(\exp(tV)) \\ &= \exp(\mu t + \frac{1}{2}\sigma^2 t^2)\frac{\Phi\left(\dfrac{\beta-\mu}{\sigma} - \sigma t\right) - \Phi\left(\dfrac{\alpha-\mu}{\sigma} - \sigma t\right)}{\Phi\left(\dfrac{\beta-\mu}{\sigma}\right) - \Phi\left(\dfrac{\alpha-\mu}{\sigma}\right)} \end{aligned} \tag{9.6}$$

The expected value of a distribution equals the first derivative of its moment generating function at $t = 0$, and hence

$$E[V] = M'(0) = \mu - \sigma\frac{\phi\left(\dfrac{\beta-\mu}{\sigma}\right) - \phi\left(\dfrac{\alpha-\mu}{\sigma}\right)}{\Phi\left(\dfrac{\beta-\mu}{\sigma}\right) - \Phi\left(\dfrac{\alpha-\mu}{\sigma}\right)} \tag{9.7}$$

Using the preceding results, we see that the components $m(\theta_0)_i$ of the vector $m(\theta_0)$ are given by

$$m(\theta_0)_i = \begin{cases} x_i^T\theta_0 + \dfrac{\phi(-x_i^T\theta_0)}{1-\Phi(-x_i^T\theta_0)}, & \text{if } y_i = 1 \\ x_i^T\theta_0 - \dfrac{\phi(-x_i^T\theta_0)}{\Phi(-x_i^T\theta_0)}, & \text{if } y_i = 0. \end{cases} \tag{9.8}$$

Formulas (9.4) and (9.8) define one step of the EM algorithm for calculating the posterior mode in probit regression. The EM algorithm for the MLE of probit regression is obtained from formulas (9.4) and (9.8) by setting R_0 as the zero matrix. (Then we need to assume that X^TX is positive definite.)

The truncated normal distribution features in many other statistical models besides the latent variable formulation of probit regression. One famous example is the tobit regression model. This is a linear regression model, where the observations are censored. Since the truncated normal distribution pops up in many different contexts, it is useful to know that there is a simple formula (9.6) for its moment generating function.

9.3 Why the EM algorithm works

The proof of the monotonicity of the EM and GEM algorithms is based on the non-negativity of the Kullback-Leibler divergence. If f and g are two densities, then the K-L divergence (or relative entropy) of g from f is defined by

$$D(f \parallel g) = \int f \, \ln \frac{f}{g}, \tag{9.9}$$

where the integral is calculated over the whole space. If the supports of f and g are not the whole space, then we use the conventions

$$f(x) \, \ln \frac{f(x)}{g(x)} = \begin{cases} 0, & \text{if } f(x) = 0, \\ \infty, & \text{if } f(x) > 0 \text{ and } g(x) = 0. \end{cases}$$

We will show that the K-L divergence is always non-negative. Therefore we can use it to measure the distance of g from f. However, the K-L divergence is not a metric (on the space of densities), since it is even not symmetric.

The proof of the non-negativity can be based on the elementary inequality

$$\ln x \le x - 1 \qquad \forall x > 0, \tag{9.10}$$

where equality holds if and only if $x = 1$. This inequality follows from the concavity of the logarithm function. The graph of a concave function lies below each of its tangents, and right hand side of (9.10) is the tangent at $x_0 = 1$.

Theorem 4. *Let f and g be densities defined on the same space. Then*

$$D(f \parallel g) \ge 0,$$

and equality holds if and only if $f = g$ (almost everywhere).

Proof. We give the proof only in the case, when f and g have the same support, i.e., when the sets $\{x : f(x) > 0\}$ and $\{x : g(x) > 0\}$ are the same (except perhaps modulo a set of measure zero). Extending the proof to handle the general case is straightforward. In the following calculation, the integral extends only over the common support of f and g.

$$\begin{aligned}(-1)D(f \parallel g) &= \int -f \ln \frac{f}{g} = \int f \ln \frac{g}{f} \\ &\leq \int f(\frac{g}{f} - 1) \qquad \text{by (9.10)} \\ &= \int (g - f) = 1 - 1 = 0.\end{aligned}$$

We have equality if and only if

$$\ln \frac{g}{f} = \frac{g}{f} - 1,$$

almost everywhere, and this happens if and only if $f = g$ almost everywhere. □

The following theorem establishes the monotonicity of EM or GEM iterations.

Theorem 5. *Define the function Q by either the equation* (9.1) *or by* (9.2). *Let θ_0 and θ_1 be any values such that*

$$Q(\theta_1 \mid \theta_0) \geq Q(\theta_0 \mid \theta_0). \tag{9.11}$$

Then, with the definition (9.1) *we have*

$$f_{\Theta|Y}(\theta_1 \mid y) \geq f_{\Theta|Y}(\theta_0 \mid y),$$

and with the definition (9.2) *we have*

$$f_{Y|\Theta}(y \mid \theta_1) \geq f_{Y|\Theta}(y \mid \theta_0).$$

In either case, if we have strict inequality in the assumption (9.11), *then we have strict inequality also in the conclusion.*

Proof. We consider first the proof for the definition (9.1). We will use the abbreviated notations, and make use of the identity

$$p(\theta \mid y) = \frac{p(\theta, z \mid y)}{p(z \mid \theta, y)}.$$

For any θ, we have

$$\begin{aligned}\ln p(\theta \mid y) &= \int p(z \mid \theta_0, y) \ln p(\theta \mid y) \, \mathrm{d}z \\ &= \int p(z \mid \theta_0, y) \ln \frac{p(\theta, z \mid y)}{p(z \mid \theta, y)} \, \mathrm{d}z \\ &= Q(\theta \mid \theta_0) - \int p(z \mid \theta_0, y) \ln p(z \mid \theta, y) \, \mathrm{d}z\end{aligned}$$

Using this identity at the points θ_1 and θ_0, we obtain

$$\begin{aligned}\ln p(\theta_1 \mid y) - \ln p(\theta_0 \mid y) & \\ &= Q(\theta_1 \mid \theta_0) - Q(\theta_0 \mid \theta_0) + \int p(z \mid \theta_0, y) \ln \frac{p(z \mid \theta_0, y)}{p(z \mid \theta_1, y)} \, \mathrm{d}z \\ &\geq Q(\theta_1 \mid \theta_0) - Q(\theta_0 \mid \theta_0),\end{aligned}$$

since the K-L divergence is non-negative. This proves the claim for (9.1).

The proof for the definition (9.2) starts from the identity

$$\begin{aligned}\ln p(y \mid \theta) &= \int p(z \mid \theta_0, y) \ln p(y \mid \theta) \, \mathrm{d}z \\ &= \int p(z \mid \theta_0, y) \ln \frac{p(y, z \mid \theta)}{p(z \mid \theta, y)} \, \mathrm{d}z \\ &= Q(\theta \mid \theta_0) - \int p(z \mid \theta_0, y) \ln p(z \mid \theta, y) \, \mathrm{d}z.\end{aligned}$$

Rest of the proof is the same as before. □

9.4 Literature

The name EM algorithm was introduced by Dempster, Laird and Rubin in [1]. Many special cases of the method had appeared in the literature already in the 1950's, but this article gave a unified structure to the previous methods. The book [3] is dedicated to the EM algorithm and its variations. Many authors have extended the EM algorithm so that one obtains also the approximate covariance matrix of the posterior, or the approximate covariance matrix of the MLE, see, e.g., [3] or [2].

Bibliography

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 45:1–38, 1977.

[2] Geof H. Givens and Jennifer A. Hoeting. *Computational Statistics*. Wiley-Interscience, 2005.

[3] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

Chapter 10

Multi-model inference

10.1 Introduction

If we consider several competing statistical models, any of which could serve as an explanation for our data, and would like to select the best of them, then we face a model selection (or a model choice, or a model comparison) problem. Instead of choosing a single best model, it might be more meaningful to combine somehow inferences obtained from all of the models, and then we may speak of model averaging. Such activities may also be called multi-model inference.

For example, in the binary regression setting with the explanatory variable x we might posit the model

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(F(\alpha + \beta x_i)), \qquad i = 1, \ldots, n,$$

where $B(p)$ is the Bernoulli distribution with success probability p, but we might want to consider several different link functions F^{-1} such as the logit, the probit and the complementary log-log transformation.

In a continuous regression problem with explanatory variable x, we might want to consider polynomials of degrees zero, one and two as the mean response,

$$\begin{aligned}
&\text{model 0:} && [Y_i \mid \alpha, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha, \sigma^2), && i = 1, \ldots, n \\
&\text{model 1:} && [Y_i \mid \alpha, \beta_1, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha + \beta_1 x_i, \sigma^2), && i = 1, \ldots, n \\
&\text{model 2:} && [Y_i \mid \alpha, \beta_1, \beta_2, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha + \beta_1 x_i + \beta_2 x_i^2, \sigma^2), && i = 1, \ldots, n.
\end{aligned}$$

One commonly occurring situation is the variable selection problem. For instance, we might want to select which of the candidate variables to use as explanatory variables in a multiple regression problem.

The usual frequentist solution to model selection in the case of nested models is to perform a series of hypothesis tests. One statistical model is said to be nested within another model, if it is a special case of the other model. In the polynomial regression example, model 0 is a special case of model 1, and model 1 is a special case of model 2. In this example a frequentist statistician would probably select among these models by using F-tests. However, one may be bothered by the fact that we actually need to make multiple tests. How should we take this into account when selecting the size of the test?

Outside the linear model framework, a frequentist statistician would compare nested models by using the asymptotic χ^2 distribution of the likelihood ratio test (LRT) statistic, but the asymptotics is valid only when the simpler model does not correspond to a parameter value at the boundary of the parameter space of the more complex model. There are important statistical models (such as the linear mixed effects model) where a natural null hypothesis corresponds to a point at the boundary of the parameter space, and then the usual χ^2 asymptotics do not apply.

In contrast to the polynomial regression example, in the binary regression example there is no natural way to nest the models, and comparing the models by hypothesis tests would be problematic.

Besides hypothesis testing, a frequentist statistician might compare models using some information criterion, such as the Akaike information criterion, AIC. This approach does not suffer from the problems we identified in the hypothesis testing approach.

In the rest of this chapter we will discuss Bayesian techniques for model selection, or more generally, to multi-model inference. The basic idea is to introduce a single encompassing model which is a union of all the alternative models. Then we use Bayes rule to derive the posterior distribution. This requires that we have successfully specified the entire collection of candidate models we want to consider. This the $\mathcal{M}$-closed case instead of the more general $\mathcal{M}$-open case, where the ultimate model collection is not known ahead of time, see [1, Ch. 6] for a deep discussion on this and other assumptions and approaches a Bayesian statistician can use in multi-model inference.

The concepts we need are borrowed from the Bayesian approach to hypothesis testing. There is no requirement that the models should be nested with respect to one another, and no problem arises if one model corresponds to a parameter value at the boundary of the parameter space of another model.

To unify the discussion we make the following conventions. The alternative models are numbered $1, \ldots, K$. The parameter vector θ_m of model m belongs to the parameter space $S_m \subset \mathbb{R}^{d_m}$. The parameter vectors $\theta_m, m = 1, \ldots, K$ of the models are considered separate: no two models share any parameters.

For example, in the binary regression example the α and β parameters for the logit link and for the probit link and for the complementary log-log link are considered separate, and we could label them, e.g., as

$$\theta_1 = (\alpha_1, \beta_1), \quad \theta_2 = (\alpha_2, \beta_2), \quad \theta_3 = (\alpha_3, \beta_3).$$

Here $S_1 = S_2 = S_3 = \mathbb{R}^2$, and $d_1 = d_2 = d_3 = 2$.

In the polynomial regression example the error variance parameters are considered separate parameters in all of the three models, the intercepts and slopes are considered separate parameters, and so on. We could label them, e.g., as

$$\theta_1 = (\alpha_0, \sigma_0^2), \quad \theta_2 = (\alpha_1, \beta_1, \sigma_1^2), \quad \theta_3 = (\alpha_2, \beta_{21}, \beta_{22}, \sigma_2^2).$$

Here $d_1 = 2$, $d_2 = 3$, $d_3 = 4$, and

$$S_1 = \mathbb{R} \times \mathbb{R}_+, \quad S_2 = \mathbb{R}^2 \times \mathbb{R}_+, \quad S_3 = \mathbb{R}^3 \times \mathbb{R}_+,$$

At first sight it may seem unnatural to separate the parameters which usually are denoted by the same symbol, such as α and σ^2 in the zeroth and the first degree polynomial regression models. To make it more acceptable, think of them in the following way.

- In the zeroth degree model α_0 is the "grand mean" and σ_0^2 is the error variance when there no explanatory variable is present in the model.
- In the first degree regression model α_1 is the intercept and σ_1^2 is the error variance when there is intercept and slope present in the model, and so on.

10.2 Marginal likelihood and Bayes factor

Handling multi-model inference in the Bayesian framework is easy, at least in principle. In the single encompassing model one needs, in addition to the parameter vectors of the different models $\theta_1, \theta_2, \ldots, \theta_K$, also a random variable M to indicate the model index. Then

$$P(M = m) \equiv p(m), \qquad m = 1, \ldots, K$$

are the prior model probabilities, which have to sum to one. Typically the prior model probabilities are chosen to be uniform. Further,

$$p(\theta_m \mid M = m) \equiv p(\theta_m \mid m),$$

is the prior on θ_m in model m,

$$p(y \mid \theta_m, M = m) \equiv p(y \mid \theta_m, m),$$

is the likelihood within model m, and

$$p(\theta_m \mid y, M = m) \equiv p(\theta_m \mid y, m)$$

is the posterior for θ_m within model m.

For model selection, the most interesting quantities are the posterior model probabilities,

$$P(M = m \mid y) \equiv p(m \mid y), \qquad m = 1, \ldots, K.$$

By Bayes rule,

$$p(m \mid y) = \frac{p(y \mid m)\, p(m)}{p(y)}, \quad \text{where} \quad p(y) = \sum_{m=1}^{K} p(y \mid m)\, p(m) \tag{10.1}$$

Here $p(y \mid m)$ is usually called the **marginal likelihood** of the data within model m, or simply the marginal likelihood of model m. Other terms like *marginal density of the data*, *integrated likelihood*, *prior predictive (density)*, *predictive likelihood* or *evidence* are also all used in the literature. The marginal likelihood of model m is obtained by averaging the likelihood using the prior as the weight, both within model m, i.e.,

$$p(y \mid m) = \int p(y, \theta_m \mid m)\, \mathrm{d}\theta_m = \int p(\theta_m \mid m)\, p(y \mid \theta_m, m)\, \mathrm{d}\theta_m. \tag{10.2}$$

In other words, the marginal likelihood is the normalizing constant needed in order to make prior times likelihood within model m to integrate to one,

$$p(\theta_m \mid y, m) = \frac{p(\theta_m \mid m)\, p(y \mid \theta_m, m)}{p(y \mid m)}.$$

The **Bayes factor** BF_{kl} for comparing model k against model l is defined to be the *ratio of posterior to prior odds*, or in more detail, the posterior odds in favor of model k against model l divided by the corresponding prior odds, i.e.,

$$\mathrm{BF}_{kl} = \frac{P(M = k \mid y)}{P(M = l \mid y)} \Big/ \frac{P(M = k)}{P(M = l)} \tag{10.3}$$

By Bayes rule (10.1), the Bayes factor equals the ratio of the two marginal likelihoods,

$$\mathrm{BF}_{kl} = \frac{p(y \mid M = k)}{p(y \mid M = l)} \tag{10.4}$$

From this we see immediately that $\mathrm{BF}_{lk} = 1/\,\mathrm{BF}_{kl}$. There are tables available (due to Jeffreys and other people) for interpreting the value of the Bayes factor.

One can compute the posterior model probabilities $p(m \mid y)$, if one knows the prior model probabilities and either the marginal likelihoods for all the models, or the Bayes factors for all pairs of models. Having done this, we may restrict our attention to the best model which has the largest posterior probability. Alternatively we might want to consider all those models whose posterior probabilities are nearly equal to that of the best model.

If one needs to form predictions for future observations Y^* which are conditionally independent of the observations, then one might form the predictions by **model averaging**, i.e., by using the predictive distribution

$$\begin{aligned} p(y^* \mid y) &= \sum_{m=1}^{K} \int p(y^*, m, \theta_m \mid y)\, \mathrm{d}\theta_m \\ &= \sum_{m=1}^{K} \int p(y^* \mid m, \theta_m, y)\, p(m \mid y)\, p(\theta_m \mid m, y)\, \mathrm{d}\theta_m \\ &= \sum_{m=1}^{K} p(m \mid y) \int p(y^* \mid m, \theta_m)\, p(\theta_m \mid m, y)\, \mathrm{d}\theta_m, \end{aligned}$$

where on the last line we used the assumption that the data Y and the future observation Y^* are conditionally independent within each of the models m, conditionally on the parameter vector θ_m. The predictive distribution for future data is obtained by averaging the within-model predictive distributions using posterior model probabilities as weights.

Similarly, we could consider the posterior distribution of a function of the parameter vector, which is meaningful in all of the candidate models. In the binary regression example, such a parameter could be LD50 (lethal dose 50 %) which is defined as the value of the covariate x which gives success probability 50 %. Such a parameter could be estimated with model averaging.

In multi-model inference one should pay close attention to the formulation of the within-model prior distributions. While the within-model posterior distributions are usually robust against the specification of the within-model prior, the same is not true for the marginal likelihood. In particular, in a multi-model situation one cannot use improper priors for the following reason. If the prior for model m is improper, i.e.,

$$p(\theta_m \mid m) \propto h_m(\theta_m)$$

where the integral of h_m is infinite, then

$$c\, h_m(\theta_m), \qquad \text{with } c > 0 \text{ arbitrary},$$

is an equally valid expression for the within-model prior. Taking $h_m(\theta_m)$ as the prior within model m in eq. (10.2) leads to the result

$$p_1(y \mid m) = \int h_m(\theta_m)\, p(y \mid \theta_m, m)\, \mathrm{d}\theta_m$$

whereas the choice $c\, h_m(\theta_m)$ leads to the result

$$p_c(y \mid m) = c\, p_1(y \mid m).$$

Therefore, if the prior for model m is improper, then we cannot assign any meaning to the marginal likelihood for model m, and the same difficulty applies to the Bayes factor, as well.

Many researchers regard the sensitivity of the marginal likelihood to the within model prior specifications a very serious drawback. This difficulty has led to many proposals for model comparison which do not depend on marginal likelihoods and Bayes factors. However, we will continue to use them for the rest of this chapter. Therefore we suppose that

- we have specified the entire collection of candidate models (this the $\mathcal{M}$-closed assumption);
- we have successfully formulated proper and informative priors for each of the candidate models.

10.3 Approximating marginal likelihoods

If we use a conjugate prior in model m, then we can calculate its marginal likelihood analytically, e.g., by using Bayes rule in the form

$$p(y \mid m) = \frac{p(\theta_m \mid m)\, p(y \mid \theta_m, m)}{p(\theta_m \mid y, m)}, \tag{10.5}$$

where θ_m is any point in the parameter space of model m, and all the terms on the right-hand side (prior density, likelihood, and posterior density, each of them within model m, respectively) are available in a conjugate situation. This form of the Bayes rule is also known by the name *candidate's formula*. In order to simplify the notation, we will drop the conditioning on the model m from the notation for the rest of this section, since we will discuss estimating the marginal likelihood for a single model at a time. For example, in the rest of this section we will write candidate's formula (10.5) in the form

$$p(y) = \frac{p(\theta)\, p(y \mid \theta)}{p(\theta \mid y)}. \tag{10.6}$$

Hopefully, leaving the model under discussion implicit in the notation does not cause too much confusion to the reader. If it does, add conditioning on m to each of the subsequent formulas and add the subscript m to each occurrence of θ and modify the text accordingly.

When the marginal likelihood is not available analytically, we may try to estimate it. One idea is based on estimating the posterior ordinate $p(\theta \mid y)$ in candidate's formula (10.6) at some point θ_h having high posterior density (such as the posterior mean estimated by MCMC). The result can be called the *candidate's estimator* for the marginal likelihood. Suppose that the parameter can be divided into two blocks $\theta = (\theta_1, \theta_2)$ such that the full conditional distributions $p(\theta_1 \mid \theta_2, y)$ and $p(\theta_2 \mid \theta_1, y)$ are both available analytically. By the multiplication rule

$$p(\theta_1, \theta_2 \mid y) = p(\theta_1 \mid y)\, p(\theta_2 \mid \theta_1, y).$$

We might estimate the marginal posterior ordinate of θ_1 at $\theta_{h,1}$ by the Rao-Blackwellized estimate

$$\hat{p}(\theta_{h,1} \mid y) = \frac{1}{N} \sum_{i=1}^{N} p(\theta_{h,1} \mid \theta_2^{(i)}, y),$$

where $(\theta_1^{(i)}, \theta_2^{(i)}), i = 1, \ldots, N$ is a sample from the posterior, e.g., produced by MCMC. Then the joint posterior at $\theta_h = (\theta_{h,1}, \theta_{h,2})$ can be estimated by

$$\hat{p}(\theta_{h,1}, \theta_{h,2} \mid y) = \hat{p}(\theta_{h,1} \mid y)\, p(\theta_{h,2} \mid \theta_{h,1}, y).$$

This approach was proposed in Chib [5] where one can also find extensions to more than two blocks.

Approximating the marginal likelihood is an ideal application for Laplace's method. Recall that the basic idea of Laplace's method is to approximate a d-dimensional integral of the form

$$I = \int g(\theta)\, \exp(L(\theta))\, \mathrm{d}\theta$$

by replacing $L(\theta)$ by its quadratic approximation centered on the mode $\tilde{\theta}$ of $L(\theta)$ and by replacing $g(\theta)$ with $g(\tilde{\theta})$. The result was

$$I \approx \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, g(\tilde{\theta})\, \mathrm{e}^{L(\tilde{\theta})},$$

where Q is the negative Hessian of $L(\theta)$ evaluated at the mode $\tilde{\theta}$.

If we start from the representation

$$p(y) = \int p(\theta)\, p(y \mid \theta)\, \mathrm{d}\theta = \int \exp\left[\log\left(p(\theta)\, p(y \mid \theta)\right)\right]\, \mathrm{d}\theta,$$

and then apply Laplace's method, we get the approximation

$$\hat{p}_{\mathrm{Lap}}(y) = p(\tilde{\theta})\, p(y \mid \tilde{\theta})\, \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}, \tag{10.7}$$

where $\tilde{\theta}$ is the posterior mode (i.e. the maximum a posterior estimate, MAP estimate), and and Q is the negative Hessian of the logarithm of the unnormalized posterior density

$$\theta \mapsto \log\left(p(\theta)\, p(y \mid \theta)\right)$$

evaluated at the mode $\tilde{\theta}$.

Another possibility is to start from the representation

$$p(y) = \int p(\theta)\, \exp\left[\log p(y \mid \theta)\right]\, \mathrm{d}\theta$$

and then integrate the quadratic approximation for the log-likelihood centered at its mode, the maximum likelihood estimate (MLE) $\hat{\theta}_{\mathrm{MLE}}$. This gives the result

$$\hat{p}_{\mathrm{Lap}}(y) = p(\hat{\theta}_{\mathrm{MLE}})\, p(y \mid \hat{\theta}_{\mathrm{MLE}})\, \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}, \tag{10.8}$$

where $Q = J(\hat{\theta}_{\mathrm{MLE}})$ is now the observed information matrix (see (6.9)) evaluated at the MLE.

One can also use various Monte Carlo approaches to approximate the marginal likelihood. Since

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, \mathrm{d}\theta,$$

naive Monte Carlo integration gives the estimate

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} p(y \mid \theta^{(i)}), \tag{10.9}$$

where we average the likelihood values using a sample $\theta^{(i)}, i = 1, \ldots, N$ from the prior $p(\theta)$. If the posterior corresponds to a large data set $y_1, \ldots, y_n$, then typically the model m likelihood is very peaked compared to the prior. In this situation the estimate (10.9) has typically huge variance, since very few of the sample points hit the region with high likelihood values, and these few values dominate the sum.

A better approach would be to write the marginal likelihood as

$$p(y) = \int \frac{p(y \mid \theta)\, p(\theta)}{g(\theta)}\, g(\theta)\, \mathrm{d}\theta,$$

where $g(\theta)$ is an importance sampling density for the model under consideration. This yields the importance sampling estimate

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{p(y \mid \theta^{(i)})\, p(\theta^{(i)})}{g(\theta^{(i)})}, \tag{10.10}$$

where $\theta^{(i)}, i = 1, \ldots, N$ is a sample drawn from the importance sampling density g. In order to obtain low variance, g should be an approximation to the posterior density, and g should have heavier tails than the true posterior. For example, g could be a multivariate t distribution centered on the posterior mode, the shape of which is chosen using an estimate of the posterior covariance matrix.

The marginal likelihood can also be estimated using an MCMC sample drawn from the posterior distribution $p(\theta \mid y)$. Let g be a probability density defined on the parameter space. Integrating the identity

$$g(\theta) = g(\theta) \frac{p(y)\, p(\theta \mid y)}{p(y \mid \theta)\, p(\theta)}$$

over the parameter space gives

$$\frac{1}{p(y)} = \int \frac{g(\theta)}{p(y \mid \theta)\, p(\theta)}\, p(\theta \mid y)\, \mathrm{d}\theta$$

If $\theta^{(i)}, i = 1, \ldots, N$ is a MCMC sample from the posterior, then we can estimate the marginal likelihood as follows,

$$\hat{p}(y) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta^{(i)})}{p(y \mid \theta^{(i)})\, p(\theta^{(i)})} \right]^{-1}. \tag{10.11}$$

Here we calculate the harmonic mean of prior times likelihood divided by the density g ordinates evaluated at the sample points, $p(y \mid \theta^{(i)})\, p(\theta^{(i)})/g(\theta^{(i)})$. This is the *generalized harmonic mean estimator* suggested by Gelfand and Dey [9]. The function g should be chosen so that it has approximately the same shape as the posterior density $p(\theta \mid y)$ but in this case the tails of g should be thin compared to the tails of the posterior.

If one selects g to be the prior $p(\theta)$ then formula (10.11) suggests that one could estimate the marginal likelihood by calculating the harmonic mean of the likelihood values $p(y \mid \theta^{(i)})$. This is the (in)famous harmonic mean estimator first discussed by Newton and Raferty [13]. The harmonic mean estimator has typically infinite variance and is numerically unstable, and therefore should not be used at all.

Besides these, many other sampling-based approaches have been proposed in the literature (e.g., bridge sampling).

After all the marginal likelihoods $p(y \mid M = j)$ have been estimated one way or another, then one can estimate the posterior model probabilities based on eq. (10.1), i.e., by using

$$\hat{p}(m \mid y) = \frac{p(m)\, \hat{p}(y \mid m)}{\sum_{j=1}^{K} p(M = j)\, \hat{p}(y \mid M = j)}, \qquad m = 1, \ldots, K.$$

The denominator is just the sum of the numerators when m takes the values from 1 to K.

An obvious way to estimate the Bayes factor BF_{kl} is to calculate the ratio of two marginal likelihood estimators,

$$\widehat{\mathrm{BF}}_{kl} = \frac{\hat{p}(y \mid M = k)}{\hat{p}(y \mid M = l)}.$$

However, there are also more direct ways of estimating the Bayes factor, such as path sampling.

10.4 BIC and other information criteria

Information criteria consist of two parts: a measure of fit of the model to the data, and a penalty for the complexity of the model. The two most famous such criteria are AIC and BIC.

Our starting point for Schwarz's Bayes(ian) Information Criterion, BIC (other acronyms: SBIC, SBC, SIC), is the Laplace approximation to the marginal posterior based on the MLE (10.8). Taking logarithms and multiplying by minus

two gives

$$-2\log p(y) \approx -2\log p(\hat{\theta}) - 2\log p(y \mid \hat{\theta}) - d\log(2\pi) + \log\det(Q).$$

where $\hat{\theta}$ is the MLE and Q is the observed information matrix (at the MLE). We concentrate on the case where we have n observations y_i which are conditionally independent, i.e.,

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta),$$

from which

$$\log p(y \mid \theta) = \sum_{i=1}^{n} \log p(y_i \mid \theta)$$

$$Q = n \left[\frac{1}{n} \sum_{i=1}^{n} (-1) \frac{\partial^2}{\partial\theta\, \partial\theta^T} \log p(y_i \mid \theta) \right]_{|\theta=\hat{\theta}}$$

It is possible to show (under general conditions) that $\hat{\theta}$ converges almost surely to some point θ_0, which delivers the best approximation to the sampling density of the data inside the parameter space of the model under consideration. Next one can argue (based on a multivariate version of the SLLN) that the average inside the square brackets above is approximately equal to the corresponding expected value $I_1(\theta_0)$, the expected (Fisher) information matrix due to a single observation, evaluated at θ_0, where

$$I_1(\theta) = -\int p(y \mid \theta) \frac{\partial^2}{\partial\theta\, \partial\theta^T} \log p(y \mid \theta)\, \mathrm{d}\theta.$$

Hence we approximate

$$Q \approx nI_1(\theta_0) \quad \Rightarrow \quad \det(Q) \approx n^d \det(I_1(\theta_0))$$

This gives

$$-2\log p(y) \approx -2\log p(y \mid \hat{\theta}) + d\log n - 2\log p(\hat{\theta}) - d\log(2\pi) + \log\det(I_1(\theta_0)).$$

The final step is to drop all the terms which remain constant as the sample size n increases, and this gives the approximation

$$-2\log p(y) \approx -2\log p(y \mid \hat{\theta}) + d\log n.$$

We have now derived the Bayesian information criterion for model m, namely

$$\mathrm{BIC}_m = -2L_m + d_m \log n. \tag{10.12}$$

Here

$$L_m = \log p(y \mid \hat{\theta}_m, m)$$

is the maximized log-likelihood for model m, d_m is the dimensionality of the model m parameter space, and n is the sample size. (Warning: in the literature you can find several mutually incompatible definitions for BIC.) This criterion can be used for rough comparison of competing models: smaller values of BIC

correspond to better models. Most of the time, more complex models lead automatically to higher values of the maximized likelihood, but the term $d_m \log n$ penalizes for increased model complexity.

The approximations involved in the derivation of BIC are rather crude, and therefore $\exp(-\frac{1}{2}\,\mathrm{BIC}_m)$ is usually a rather poor approximation to the marginal likelihood of model m. Nevertheless, if we approximate $p(y \mid m)$ by $\exp(-\frac{1}{2}\,\mathrm{BIC}_m)$, and assume that the prior model probabilities are equal, then we may estimate the posterior model probabilities by

$$\hat{p}(m \mid y) = \frac{\exp(-\frac{1}{2}\,\mathrm{BIC}_m)}{\sum_{k=1}^{K} \exp(-\frac{1}{2}\,\mathrm{BIC}_k)}. \tag{10.13}$$

BIC resembles the equally famous Akaike information criterion, AIC,

$$\mathrm{AIC}_m = -2L_m + 2d_m.$$

As for BIC, also for AIC smaller is better. In addition, the alphabet soup of information criteria includes such acronyms as AIC_c (corrected AIC), cAIC (conditional AIC), mAIC; AFIC; BFIC; DIC; FIC; HQ; NIC; QAIC and QAIC_c; RIC; TIC; WIC. Furthermore, there are several other famous model selection criteria available, such as Mallows' C_p (for regression problems with normal errors), or Akaike's FPE (final prediction error). Also Rissanen's MDL (minimum description length) principle can be used. See, e.g., Burnham and Anderson [2] and Claeskens and Hjort [6].

In some statistical models it is not always clear what one should use as the sample size n in these information criteria. What is more, in complex models the number of parameters is not necessarily clearly defined. Spiegelhalter *et al.* [15] suggest that in such a situation one may use their deviance information criterion, DIC, defined by

$$\mathrm{DIC}_m = 2\overline{D(\theta_m, m)} - D(\bar{\theta}_m, m), \tag{10.14}$$

where $D(\theta_m, m)$ is the deviance, or minus twice the log-likelihood of model m,

$$D(\theta_m, m) = -2 \log p(y \mid \theta_m, m),$$

$\bar{\theta}_m$ is the posterior mean of θ_m, and $\overline{D(\theta_m, m)}$ is the posterior mean of $D(\theta_m, m)$ within model m. These quantities are estimated using separate MCMC runs for each of the models. WinBUGS and OpenBUGS have automatic facilities for calculating DIC, and therefore it has become widely used among Bayesian statisticians. As with AIC and BIC, smaller DIC indicates a better model.

The authors interpret

$$d_m^{\mathrm{eff}} = \overline{D(\theta_m, m)} - D(\bar{\theta}_m, m)$$

as the number of effective parameters for model m, and therefore DIC_m can written in the form

$$\mathrm{DIC}_m = D(\bar{\theta}_m, m) + 2d_m^{\mathrm{eff}},$$

which shows its connection to AIC. The authors show that d_m^{eff} gives a reasonable definition for the effective number of parameters in many cases. If there is strong conflict between the prior and the data, then the effective number of parameters may turn out have a negative value, which does not make sense.

In order to use DIC, one must decide which expression to use as the likelihood. In complex statistical models, e.g., hierarchical models or random effects models, even this choice is not clear cut. Consider the hierarchical model, which has a prior on the hyperparameters ψ and which factorizes as follows

$$p(y, \theta, \psi) = p(y \mid \theta)\, p(\theta \mid \psi)\, p(\psi).$$

If one focuses the attention to the parameter vector θ, then the likelihood expression is $p(y \mid \theta)$. However, it would be equally valid to consider the vector ψ to be the true parameter vector. If one focuses on ψ, then one should select

$$p(y \mid \psi) = \int p(y, \theta \mid \psi)\, \mathrm{d}\psi = \int p(y \mid \theta)\, p(\theta \mid \psi)\, \mathrm{d}\psi$$

as the likelihood. In some models $p(y \mid \psi)$ is available in closed form. Otherwise, evaluating this likelihood may be problematic. Generally, the DIC values for $p(y \mid \theta)$ and $p(y \mid \psi)$ are different. Spiegelhalter *et al.* suggest that one should formulate clearly the focus of the analysis, and calculate DIC using the corresponding likelihood expression. They also point out that DIC_m changes, if one reparametrizes model m.

10.5 Sum space versus product space

In this section we discuss an embedding of the multi-model inference problem in the product-space formulation of the problem. We revert to the explicit notation of Section 10.2. Let

$$S_m \subset \mathbb{R}^{d_m}, \qquad m = 1, \ldots, K$$

be the parameter space of model m. We call the set

$$S_{\mathrm{sum}} = \cup_{m=1}^{K} \{m\} \times S_m \tag{10.15}$$

the sum of the parameter spaces. (In topology, this would be called the topological sum, direct sum, disjoint union or coproduct of the spaces S_m.) Any point $x \in S_{\mathrm{sum}}$ is of the form

$$x = (m, \theta_m), \qquad \text{where } m \in \{1, \ldots, K\} \text{ and } \theta_m \in S_m.$$

The quantities of inferential interest discussed in Section 10.2 can be defined based on the joint posterior

$$p(m, \theta_m \mid y), \qquad m \in \{1, \ldots, K\}, \quad \theta_m \in S_m,$$

which itself is defined on the sum space through the joint distribution specification

$$p(m, \theta_m, y) = p(m)\, p(\theta_m \mid m)\, p(y \mid \theta_m, m), \qquad m \in \{1, \ldots, K\}, \quad \theta_m \in S_m.$$

Designing a MCMC algorithm which uses the sum space as its state space is challenging. For instance, the dimensionality of the parameter vector may change each time the model indicator changes. Specifying the sum-space formulation directly in BUGS seems to be impossible, since in the sum-space formulation parameter θ_m exits only when the model indicator has the value m.

Green [11] was first to propose a trans-dimensional MCMC algorithm which works directly in the sum space, and called it the reversible jump MCMC (RJMCMC) algorithm.

Most of the other multi-model MCMC algorithms are conceptually based on the product-space formulation, where the state space is the Cartesian product of the model space $\{1, \ldots, K\}$ and the Cartesian product of the parameter spaces of the models,

$$S_{\text{prod}} = S_1 \times S_2 \times \cdots \times S_K. \tag{10.16}$$

For the rest of the section, θ without a subscript will denote a point point $\theta \in S_{\text{prod}}$. It is of the form

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K), \tag{10.17}$$

where each of the $\theta_m \in S_m$. The product space is larger than the sum space, and the product-space formulation requires that we set up the joint distribution

$$p(m, \theta, y), \qquad m \in \{1, \ldots, K\}, \quad \theta \in S_{\text{prod}}.$$

In contrast, in the sum-space formulation the parameters $\{\theta_k, k \neq m\}$ do not exist on the event $M = m$, and so we cannot speak of

$$p(m, \theta, y) = p(m, \theta_1, \ldots, \theta_K, y)$$

within the sum-space formulation. We are obliged to set up the product-space formulation in such a way that the marginals

$$p(m, \theta_m, y), \qquad m \in \{1, \ldots, K\}$$

remain the same as in the original sum-space formulation. For this reason we will not make a notational difference between the sum-space and the product-space formulation of the multi-model inference problem.

The preceding means that we embed the multi-model inference problem in the product-space formulation. While specifying the sum-space model is not possible in WinBUGS/OpenBUGS, it is straightforward to specify the product-space version of the same problem.

When we do posterior inference in the product-space formulation, only the marginals

$$p(m, \theta_m \mid y), \qquad m \in \{1, \ldots, K\}$$

of the joint posterior

$$p(m, \theta \mid y) = p(m, \theta_1, \ldots, \theta_K \mid y)$$

are of inferential relevance. The other aspects of the joint distribution are only devices, which allow us to work with the easiear product-space formulation.

If $(m^{(i)}, \theta^{(i)}), i = 1, \ldots, N$ is a sample from the posterior $p(m, \theta \mid y)$, then for inference we use only the component $\theta^{(i)}_{m^{(i)}}$ of $\theta^{(i)}$, which is the parameter vector of that model $m^{(i)}$ which was visited during the i'th iteration. In particular, the posterior model probabilities $p(M = j \mid y)$ can be estimated by tabulating the relative frequencies of each of the possibilities $m^{(i)} = j$.

10.6 Method of Carlin and Chib

Carlin and Chib [3] use the product-space formulation, where

$$p(m, \theta, y) = p(m)\, p(\theta, y \mid m), \tag{10.18}$$

and $p(m)$ is the familiar model m prior probability. The conditional density $p(\theta, y \mid m)$ is selected to be

$$p(\theta, y \mid m) = p(\theta_m \mid m)\, p(y \mid \theta_m, m) \prod_{k \neq m} g_k(\theta_k \mid y) \tag{10.19}$$

Here $p(\theta_m \mid m)$ and $p(y \mid \theta_m, m)$ are the prior and the likelihood within model m, respectively. In addition, we need K densities $g_k(\theta_k \mid y)$, $k = 1, \ldots, K$ which can be called *pseudo priors* or *linking densities*. The linking density $g_k(\theta_k \mid y)$ is an arbitrary density on the parameter space of model k. It can be shown that this is a valid formulation of the product-space joint density. No circularity results from allowing the linking densities to depend on the data. Further, this specification leads to the marginals $p(m, \theta_m, y)$ of the sum-space formulation irrespective of how one specifies the linking densities.

Let us consider the case of two models ($K = 2$) in more detail. According to (10.18) and (10.19), the joint density $p(m, \theta, y)$ is

$$\begin{cases} p(M = 1)\, p(\theta_1 \mid M = 1)\, p(y \mid \theta_1, M = 1)\, g_2(\theta_2 \mid y) & \text{when } m = 1 \\ p(M = 2)\, p(\theta_2 \mid M = 2)\, p(y \mid \theta_2, M = 2)\, g_1(\theta_1 \mid y) & \text{when } m = 2. \end{cases}$$

We see easily that the marginal densities $p(m, \theta_m, y), m = 1, 2$ are the same as in the sum-space formulation: just integrate out

$$\begin{aligned} \theta_2 \quad & \text{from } p(m = 1, \theta_1, \theta_2, y) \\ \theta_1 \quad & \text{from } p(m = 2, \theta_1, \theta_2, y). \end{aligned}$$

Hence we have checked the validity of the specification.

While the specification of the linking densities $g_k(\theta_k \mid y)$ does not influence the validity of the product-space formulation, this matter does have a critical influence on the efficiency of the ensuing MCMC algorithm. A recommended choice is to select $g_k(\theta_k \mid y)$ to be a tractable approximation to the posterior distribution within model k, such as a multivariate normal approximation or a multivariate t approximation. Building such approximations usually requires pilot MCMC runs of all the models under consideration.

Carlin and Chib use the Gibbs sampler. For this we need the full conditionals. First,

$$p(m \mid \theta, y) \propto p(m, \theta, y), \qquad m = 1, \ldots, K.$$

which is easy to simulate since it is a discrete distribution. Next,

$$p(\theta_m \mid M = m, \theta_{-m}, y) \propto p(\theta_m \mid M = m)\, p(y \mid \theta_m, M = m).$$

Hence this full conditional is the within model m posterior distribution. Finally, for $k \neq m$

$$p(\theta_k \mid M = m, \theta_{-k}, y) = g_k(\theta_k \mid y)$$

is the linking density for θ_k.

These full conditionals lead to a Gibbs sampler (or a Metropolis-within-Gibbs sampler), where one first selects a new value m^{cur} for the model indicator, drawing the new value from the full conditional $p(m \mid \theta, y)$. After this, one updates the parameter vectors of all the models. For m equal to m^{cur} (for the currently visited model), the new value for θ_m is drawn from the posterior of model m (and if this is not feasible, one may execute a M–H step for the same target $p(\theta_m \mid y, m)$, instead). For all other values of k, the new value of θ_k is drawn from the linking density $g_k(\theta_k \mid y)$.

Many other product-space algorithms have been developed as well, see [10] for a review.

10.7 Reversible jump MCMC

Green's reversible jump MCMC algorithm (RJMCMC) [11] uses a Markov chain whose state space is the sum space. We discuss a simplified version of RJMCMC, where there is only one type of move available for moving from model m to model k. We also assume that the distributions of the parameter vectors θ_m in all of the models are continuous.

The RJMCMC works like the Metropolis–Hastings algorithm. One first proposes a new state, and then accepts the proposed state as the new state of the Markov chain, if $v < r$, where r is the test ratio and v is a fresh uniform $\text{Uni}(0, 1)$ random variate. The difference lies in the details: how the proposed state is generated, and how the test ratio is calculated. The state space of the Markov chain is the sum space S_{sum}, and the target distribution π is the posterior distribution

$$\pi(m, \theta_m) = p(m, \theta_m \mid y), \qquad m \in \{1, \ldots, K\}, \quad \theta_m \in S_m.$$

When the current state of the chain is (m, θ_m), then the proposal (k, θ_k) and the test ratio r are calculated as described in algorithm 20. The proposed model k is drawn from the pmf $\beta(\cdot \mid m)$. If $k = m$, then one executes an ordinary M–H step within model m. If $k \neq m$, then one proposes a new parameter vector θ_k in model k as follows. First one generates a noise vector u_m associated with θ_m from noise density $g(\cdot \mid \theta_m, m \to k)$ specific for the move $m \to k$. Then one calculates θ_k and u_k by applying the so called dimension-matching function $T_{m\to k}$. The dimension-matching functions are defined for all moves $m \neq k$, and they have to satisfy the following compatibility conditions, which are also called dimension-matching conditions.

We assume that for each move $m \to k$ where $m \neq k$ there exists a diffeomorphic correspondence

$$(\theta_k, u_k) = T_{m\to k}(\theta_m, u_m)$$

with inverse $T_{k\to m}$, i.e.,

$$(\theta_k, u_k) = T_{m\to k}(\theta_m, u_m) \quad \Leftrightarrow \quad (\theta_m, u_m) = T_{k\to m}(\theta_k, u_k). \tag{10.20}$$

Here u_m is the noise variable associated with θ_m and u_k is the noise variable associated with θ_k (for the move $m \to k$). Here the dimensions have to match,

$$\dim(\theta_m) + \dim(u_m) = \dim(\theta_k) + \dim(u_k),$$

Algorithm 20: One step of the RJMCMC algorithm.

Input: The current state of the chain is (m, θ_m).
Assumption: The correspondences (10.20) are diffeomorphic.
Result: Proposed next value (k, θ_k) as well as the test ratio r.

Draw k from the pmf $\beta(k \mid m)$.
if $k = m$ **then**
    generate the proposal θ_k with some M–H proposal mechanism within model m, and calculate r with the ordinary formula for the M–H ratio.
else
    Draw the noise variable u_m from density $g(u_m \mid \theta_m, m \to k)$. (This step is omitted, if the move $m \to k$ is deterministic.)
    Calculate θ_k and u_k by the diffeomorphic correspondence specific for the move $m \to k$,

$$(\theta_k, u_k) \leftarrow T_{m\to k}(\theta_m, u_m).$$

    Calculate r by

$$r \leftarrow \frac{\pi(k, \theta_k)}{\pi(m, \theta_m)} \frac{\beta(m \mid k)}{\beta(k \mid m)} \frac{g(u_k \mid \theta_k, k \to m)}{g(u_m \mid \theta_m, m \to k)} \left| \frac{\partial(\theta_k, u_k)}{\partial(\theta_m, u_m)} \right|$$

end

since otherwise such a diffeomorphism cannot exist.

Notice the following points concerning this method.

- When we calculate the test ratio r for the move $m \to k$, we have to use the quantities $\beta(m \mid k)$ and $g(u_k \mid \theta_k, k \to m)$ which correspond to the distributions from which we simulate, when the current state is (k, θ_k) and the move is selected to be $k \to m$.

- The Jacobian is the Jacobian of the transformation which maps (θ_m, u_m) to (θ_k, u_k), when the move is $m \to k$, i.e.,

$$\frac{\partial(\theta_k, u_k)}{\partial(\theta_m, u_m)} = \frac{\partial T_{m\to k}(\theta_m, u_m)}{\partial(\theta_m, u_m)}.$$

  We will see in Sec. 11.8 that the Jacobian term arises from the change-of-variables formula for integrals, the reason being the fact that the proposal θ_k is calculated in an indirect way, by applying the deterministic function $T_{m\to k}$ to the pair (θ_m, u_m).

- One of the moves $m \to k$ or $k \to m$ can deterministic. If the move $m \to k$ is deterministic, then the associated noise variable, u_m is not defined nor simulated, the dimension-matching function is $(\theta_k, u_k) = T_{m\to k}(\theta_m)$, and the noise density value, $g(u_m \mid \theta_m, m \to k)$ gets replaced by the constant one. The same rules apply, when the move $k \to m$ is deterministic.

- The target density ratio is calculated by

$$\frac{\pi(k, \theta_k)}{\pi(m, \theta_m)} = \frac{P(M = k)}{P(M = m)} \frac{p(\theta_k \mid M = k)}{p(\theta_m \mid M = m)} \frac{p(y \mid M = k, \theta_k)}{p(y \mid M = m, \theta_m)}$$

- The test ratio r can be described verbally as

$$r = \text{(prior ratio)} \times \text{(likelihood ratio)} \times \text{(proposal ratio)} \times \text{(Jacobian)}$$

It is possible to extend the method to the situation where we have discrete components in the state vectors θ_m of some of the models m. It is also possible to have more than one type of move between any given models. See the original paper by Green [11] for more details. The choice of the dimension-matching functions is critical to ensure good mixing of the Markov chain. In this respect, Green's automatic generic trans-dimensional sampler [12] seems to be very promising.

10.8 Discussion

In this chapter we have seen many different approaches for estimating the posterior model probabilities, which are central quantities both for model selection and model averaging. One approach is to estimate the marginal likelihoods for all of the models, and a distinct approach is to set up an MCMC algorithm which works over the model space and the parameter spaces of each of the models. Many variations are possible within each of the two approaches. What are the pros and cons of these approaches?

If the list of candidate models is short, then it is usually easy to estimate the marginal likelihoods for each of the models separately. However, if the list of candidate models is large and if it is suspected that only few of the models are supported by the data, then the best option might be to implement a multi-model MCMC sampler. However, getting the multi-model sampler to mix across the different models can be a challenging exercise and might require investigating pilot runs within each of the candidate models. Mixing within the parameter space of a single model is usually very much easier to achieve.

10.9 Literature

In addition to the original articles, see the books [4, 14, 7, 8],which also address model checking (model assessment, model criticism) which we have neglected in this chapter.

Bibliography

[1] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* John Wiley & Sons, 2000. First published in 1994.

[2] Kenneth B. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer, 2nd edition, 2002.

[3] Bradley P. Carlin and Siddhartha Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57:473–484, 1995.

[4] Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 3rd edition, 2009.

[5] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[6] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[7] Peter Congdon. *Bayesian Statistical Modelling*. Wiley, 2nd edition, 2006.

[8] Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, second edition, 2006.

[9] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, 56:501–514, 1994.

[10] Simon J. Godsill. On the relationship between Markov chain Monte Carlo mthods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10:230–248, 2001.

[11] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[12] Peter J. Green. Trans-dimensional Markov chain Monte Carlo. In Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Pres, 2003.

[13] M. A. Newton and A. E. Raferty. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.

[14] Ioannis Ntzoufras. *Bayesian Modeling Using WinBUGS*. Wiley, 2009.

[15] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64:583–639, 2002.

Chapter 11

MCMC theory

In this chapter we will finally justify the usual MCMC algorithms theoretically using the machinery of general state space Markov chains. We will prove that the Markov chains corresponding to our MCMC algorithms have the correct invariant distributions, using the concept of reversibility of a Markov chain. Additionally, we will try to understand, what the concept of irreducibility of a Markov chain means and also touch on the topic of Markov chain central limit theorems.

11.1 Transition kernel

Let S be the state space of a homogeneous Markov chain

$$\Theta^{(0)}, \Theta^{(1)}, \Theta^{(2)}, \dots$$

This means that each of the RVs $\Theta^{(i)}$ takes values in the space S. S is usually some subset of the Euclidean space. When the chain corresponds to a MCMC algorithm, where the support of the target distribution is not the whole space under consideration, then we usually choose S equal to the support of the target distribution.

Let $K(\theta, A)$ be the transition (probability) kernel of the homogeneous Markov chain, i.e., we suppose that for all $A \subset S$ we have

$$K(\theta, A) = P(\Theta^{(t+1)} \in A \mid \Theta^{(t)} = \theta). \tag{11.1}$$

As a function of $A \subset S$, the transition kernel $K(\theta, A)$ is the conditional distribution of $\Theta^{(t+1)}$ given that $\Theta^{(t)} = \theta$. Of course,

$$K(\theta, S) = 1 \qquad \forall \theta.$$

If μ is the initial distribution of the chain, i.e.,

$$\mu(A) = P(\Theta^{(0)} \in A), \qquad A \subset S,$$

then the joint distribution of $\Theta^{(0)}$ and $\Theta^{(1)}$ is

$$P_\mu(\Theta^{(0)} \in A, \Theta^{(1)} \in B) = \int_A \mu(\mathrm{d}\theta_0) K(\theta_0, B).$$

Hence the distribution of the next state is

$$P_\mu(\Theta^{(1)} \in B) = \int \mu(\mathrm{d}\theta)\, K(\theta, B), \qquad B \subset S. \tag{11.2}$$

When the domain of integration is not indicated, as here, the integral is taken over the whole space S. Here the integral is the Lebesgue integral of the function $\theta \mapsto K(\theta, B)$ with respect to the measure μ. We write the initial distribution itself, or its density, as a subscript to the P-symbol, if need be.

Recall that we call $\pi(\theta)$ a density even if it represents a discrete distribution with respect to some components of θ and a continuous distribution for others. Then integrals involving the density $\pi(\theta)$ can actually be sums with respect to some components of θ and integrals with respect to the others. If the initial distribution has a density $\pi(\theta)$, then the initial distribution itself is given by

$$\mu(A) = \int_A \pi(\theta)\, \mathrm{d}\theta.$$

In that case, the distribution of the next state given in (11.2) can be written as

$$P_\mu(\Theta^{(1)} \in B) = \int \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta \qquad B \subset S. \tag{11.3}$$

However, this distribution may or may not admit a density; which case obtains depends on the nature of the transition kernel.

In some cases (but not always) the transition kernel can be obtained from a transition density $k(\theta_1 \mid \theta_0)$ by integration,

$$K(\theta_0, B) = \int_B k(\theta_1 \mid \theta_0)\, \mathrm{d}\theta_1.$$

In such a case $k(\theta_1 \mid \theta_0)$ is the conditional density of $\Theta^{(1)}$ conditionally on $\Theta^{(0)} = \theta_0$. If the initial distribution has the density π, then (11.3) can be written as

$$P_\pi(\Theta^{(1)} \in B) = \int_{\theta_1 \in B} \int \pi(\theta_0)\, k(\theta_1 \mid \theta_0)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_0.$$

That is, the density of $\Theta^{(1)}$ can be obtained from the joint density $\pi(\theta_0)\, k(\theta_1 \mid \theta_0)$ by marginalization.

The joint distribution of the states $\Theta^{(0)}, \Theta^{(1)}$ and $\Theta^{(2)}$ is determined by

$$P_\mu(\Theta^{(0)} \in A_0, \Theta^{(1)} \in A_1, \Theta^{(2)} \in A_2) = \int_{\theta_0 \in A_0} \int_{\theta_1 \in A_1} \mu(\mathrm{d}\theta_0)\, K(\theta_0, \mathrm{d}\theta_1)\, K(\theta_1, A_2)$$

where μ is the initial distribution. If the initial distribution has density π, and the transition kernel can be obtained from transition density $k(\theta_1 \mid \theta_0)$, then the previous formula just states that the joint density of $\Theta^{(0)}$, $\Theta^{(1)}$ and $\Theta^{(2)}$ is

$$\pi(\theta_0)\, k(\theta_1 \mid \theta_0)\, k(\theta_2 \mid \theta_1).$$

Iterating, we see that the initial distribution μ and the transition kernel K together determine the distribution of the homogeneous Markov chain.

11.2 Invariant distribution and reversibility

The density $\pi(\theta)$ is an **invariant density** (or stationary density or equilibrium density) of the chain (or of its transition kernel), if the Markov chain preserves it in the following sense. When the initial state has the invariant distribution corresponding to the invariant density, then all the consecutive states have to have the same invariant distribution. In particular, when the initial distribution has the invariant density π, then the the distribution of $\Theta^{(1)}$ also has to have the density π. That is,

$$P_\pi(\Theta^{(0)} \in B) = P_\pi(\Theta^{(1)} \in B), \qquad \forall B \subset S. \tag{11.4}$$

If this holds, then by induction also all the consecutive states have the same invariant distribution, so this requirement is equivalent with the requirement that π is the invariant density of the Markov chain.

By (11.3), the requirement (11.4) can also be written in terms of the transition kernel,

$$\int_B \pi(\theta)\, \mathrm{d}\theta = \int \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta, \qquad \forall B \subset S. \tag{11.5}$$

A given transition kernel may have more than one invariant densities. E.g., the kernel

$$K(\theta, A) = 1_A(\theta), \qquad A \subset S$$

corresponds to the Markov chain which stays for ever at the same state where it starts. Obviously, any probability distribution is an invariant distribution for this trivial chain. Staying put obviously preserves any target distribution, but at the same time, this is obviously useless for the purpose of exploring the target. Useful Markov chains are ergodic, and then the invariant density can be shown to be unique.

One simple way to ensuring that a Markov chain has a specified invariant density π is to construct the transition kernel K so that it is **reversible** with respect to π. This means that the condition

$$P_\pi(\Theta^{(0)} \in A, \Theta^{(1)} \in B) = P_\pi(\Theta^{(0)} \in B, \Theta^{(1)} \in A) \tag{11.6}$$

holds for every $A, B \subset S$. This means that

$$(\Theta^{(0)}, \Theta^{(1)}) \stackrel{\mathrm{d}}{=} (\Theta^{(1)}, \Theta^{(0)}), \qquad \text{when} \quad \Theta^{(0)} \sim \pi,$$

that is, the joint distribution of the pair $(\Theta^{(0)}, \Theta^{(1)})$ is the same as the joint distribution of the pair $(\Theta^{(1)}, \Theta^{(0)})$, when the chain is started from the invariant distribution. Of course, the same result then extends to all pairs $(\Theta^{(i)}, \Theta^{(i+1)})$, where $i \geq 0$.

Expressed in terms of the transition kernel, the condition (11.6) for reversibility becomes

$$\int_A \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta = \int_B \pi(\phi)\, K(\phi, A)\, \mathrm{d}\phi, \qquad \forall A, B \subset S. \tag{11.7}$$

These equations are also called the **detailed balance** equations.

Theorem 6. *If the transition kernel K is reversible for π, then π is an invariant density for the chain.*

Proof. For any $A \subset S$

$$P_\pi(\Theta^{(0)} \in A) = P_\pi(\Theta^{(0)} \in A, \Theta^{(1)} \in S) = P_\pi(\Theta^{(0)} \in S, \Theta^{(1)} \in A)$$
$$= P_\pi(\Theta^{(1)} \in A). \qquad \square$$

11.3 Finite state space

It is instructive to specialize the preceding concepts for the case of a finite state space, which may be more familiar to the reader. Consider a Markov chain on the finite state space

$$S = \{1, \ldots, k\}.$$

Now we can identify the transition kernel with the transition matrix $P = (p_{ij})$ with entries

$$p_{ij} = P(\Theta^{(t+1)} = j \mid \Theta^{(t)} = i), \qquad i = 1, \ldots, k, \quad j = 1, \ldots, k.$$

It is customary to let the first index denote the present state, and the second index the possible values of the next state.

The entries of the transition matrix have obviously the following properties,

$$p_{ij} \geq 0 \quad \forall i, j; \qquad \sum_{j=1}^{k} p_{ij} = 1, \quad \forall i.$$

All the elements are non-negative and all the rows sum to one. Such a matrix is called a stochastic matrix. The transition kernel corresponding to the transition matrix is

$$K(i, A) = \sum_{j \in \{1,\ldots,k\} \cap A} p_{ij}.$$

If the pmf of the initial distribution is expressed as the row vector $\pi^T = [\pi_1, \ldots, \pi_k]$, then the pmf at time one is

$$\sum_i \pi_i p_{ij} = [\pi^T P]_j,$$

i.e., it is the j'th entry of the row vector $\pi^T P$.

The probability row vector $\pi^T = [\pi_1, \ldots, \pi_k]$ is stationary if and only if

$$\pi^T = \pi^T P,$$

which means that π^T has to be a left eigenvector of P corresponding to eigenvalue one, and π has to be a probability vector: its entries must be non-negative and sum to one. (A left eigenvector of P is simply the transpose of an ordinary eigenvector [or right eigenvector] of P^T).

In a finite state space the transition matrix P is reversible with respect to π, if

$$\pi_i \, p_{ij} = \pi_j \, p_{ji}, \quad \forall i, j.$$

Then π is an invariant pmf, since for any j

$$\sum_i \pi_i \, p_{ij} = \sum_i \pi_j \, p_{ji} = \pi_j \sum_i p_{ji} = \pi_j.$$

11.4 Combining kernels

A simulation algorithm, where one calculates the new state θ' based on the old state θ and some freshly generated random numbers corresponds to the kernel $K(\theta, A)$, where

$$K(\theta, A) = P(\Theta' \in A \mid \theta).$$

Now suppose that we have two simulation codes, which correspond to two different kernels $K_1(\theta, A)$ and $K_2(\theta, A)$. What is the transition kernel from θ to θ'', if we first calculate θ' by the code corresponding to $K_1(\theta, \cdot)$, and then calculate θ'' using the code corresponding to $K_2(\theta', \cdot)$? Notice that in the second step the initial value is the state where we ended up after the first step. The new piece of code corresponds to a transition kernel which we will denote by

$$K_1 K_2.$$

This is can be called the cycle of K_1 and K_2. In a finite state space $K_1 K_2$ corresponds to multiplying the transition matrices P_1 and P_2 to form the transition matrix $P_1 P_2$.

If we have d kernels $K_1, \ldots, K_d$, then we can define the **cycle** of the kernels $K_1, \ldots, K_d$ by

$$K_1 K_2 \cdots K_d,$$

which corresponds to executing the simulations corresponding to the kernels sequentially, always starting from the state where the previous step took us. If K_j is the transition kernel of the jth component Gibbs updating step, then the combined kernel $K_1 \cdots K_d$ is the kernel the deterministic scan Gibbs sampler, where the updates are carried out in the order $1, 2, \ldots, d$.

Now suppose that π is an invariant density for all kernels K_j. If the initial state Θ has the density π, then after drawing Θ' from the kernel $K_1(\theta, \cdot)$, the density of Θ' is π. When we then simulate Θ'' from the kernel $K_2(\theta', \cdot)$, its density is again π, and so on. Therefore the cycle kernel

$$K_1 K_2 \cdots K_d$$

also has π as its invariant density.

Now suppose that we have d transition kernels K_j. Suppose also that $\beta_1, \ldots, \beta_d$ is a probability vector. Then the kernel

$$K(\theta, A) = \sum_{j=1}^{d} \beta_j K_j(\theta, A)$$

is called a **mixture** of the kernels $K_1, \ldots, K_d$. It corresponds to the following simulation procedure. We draw j from the pmf $\beta_1, \ldots, \beta_d$ and then draw the new value θ' using the kernel $K_j(\theta, \cdot)$. If K_j is the jth updating step of a Gibbs sampler, then K is the transition kernel of the random scan Gibbs sampler corresponding to selecting the component to be updated using the probabilities $\beta_1, \ldots, \beta_d$.

Suppose that all the kernels K_j have π as an invariant density. Then also the mixture $K = \sum \beta_j K_j$ has the same invariant density, since

$$\int_A \pi(\theta)\, \mathrm{d}\theta = \int \pi(\theta)\, K_j(\theta, A)\, \mathrm{d}\theta, \qquad \forall j \quad \forall A \subset S,$$

and hence

$$\int_A \pi(\theta)\,\mathrm{d}\theta = \sum_{j=1}^{d} \beta_j \int_A \pi(\theta)\,\mathrm{d}\theta = \sum_{j=1}^{d} \beta_j \int \pi(\theta)\,K_j(\theta, A)\,\mathrm{d}\theta = \int \pi(\theta)\,K(\theta, A)\,\mathrm{d}\theta.$$

For this argument to work, it is critical that the mixing vector $\beta_1, \ldots, \beta_d$ does not depend on the present state θ.

We have proved the following theorem.

Theorem 7. *If π is an invariant density for each of the kernels $K_1, \ldots, K_d$, then it is also an invariant density for the cycle kernel $K_1 \cdots K_d$.*

If π is an invariant density for each of the kernels $K_1, \ldots, K_d$ and $\beta_1, \ldots, \beta_d$ is a probability vector, i.e., each $\beta_i \geq 0$ and $\beta_1 + \cdots + \beta_d = 1$, then π is also an invariant density for the mixture kernel $\sum_{j=1}^{d} \beta_j K_j$.

11.5 Invariance of the Gibbs sampler

Suppose that the target density is $\pi(\theta)$, where θ is divided into components

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d).$$

Now consider the transition kernel K_j corresponding to the j**th component Gibbs sampler**. This sampler updates the jth component θ_j of θ only and keeps all the other components θ_{-j} at their original values. The sampler draws a new value θ_j' for θ_j from the corresponding full conditional density, which we denote by

$$\pi_j(\theta_j \mid \theta_{-j}).$$

A key observation is the identity

$$\pi(\theta) = \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j}),$$

where $\pi(\theta_{-j})$ is the marginal density of all the other components except θ_j.

Theorem 8. *The transition kernel corresponding to the jth component Gibbs sampler has π as its invariant density.*

Proof. Let the initial state Θ have density π, and let Θ_j' be drawn from the jth full conditional density. Then the joint distribution of Θ and Θ_j' has the density

$$\pi(\theta)\,\pi_j(\theta_j' \mid \theta_{-j}) = \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j})\,\pi_j(\theta_j' \mid \theta_{-j}).$$

After the update, the state is (Θ_j', Θ_{-j}). We obtain its density by integrating out the variable θ_j from the joint density of Θ and Θ_j', but

$$\begin{aligned}\int \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j})\,\pi_j(\theta_j' \mid \theta_{-j})\,\mathrm{d}\theta_j &= \pi_j(\theta_j' \mid \theta_{-j})\,\pi(\theta_{-j}) \int \pi(\theta_j \mid \theta_{-j})\,\mathrm{d}\theta_j \\ &= \pi_j(\theta_j' \mid \theta_{-j})\,\pi(\theta_{-j}) = \pi(\theta').\end{aligned}$$

Therefore the updated state has the density π. □

It now follows from theorem 7 that the systematic scan and the random scan Gibbs samplers have π as their invariant distribution.

It can also be shown that the transition kernel K_j of the jth Gibbs update is reversible with respect to π. From this it follows that the transition kernel $\sum_j \beta_j K_j$ of the random scan Gibbs sampler is also reversible with respect to π. However, the transition kernel of the systematic scan Gibbs sampler is not usually reversible. (The distinction between reversible and non-reversible kernels makes a difference when one discusses the regularity conditions needed for the Markov chain central limit theorems.)

11.6 Reversibility of the M–H algorithm

Proving that the Metropolis–Hastings update leaves the target density invariant requires more effort than proving the same property for the Gibbs sampler.

Let the initial state Θ be θ and let the next state be denoted by Φ. Recall that Φ is obtained from θ by the following steps.

- We generate the proposal Θ' from the proposal density $q(\theta' \mid \theta)$, and independently $U \sim \text{Uni}(0, 1)$.
- We set
$$\Phi = \begin{cases} \Theta', & \text{if } U < r(\theta, \Theta') \text{ (accept)} \\ \theta, & \text{otherwise (reject),} \end{cases}$$
where the M–H ratio $r(\theta, \theta')$ is defined by
$$r(\theta, \theta') = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} \tag{11.8}$$

Notice that $r(\theta, \theta')$ can be greater than one, and hence the probability of acceptance, conditionally on $\Theta = \theta$ and $\Theta' = \theta'$ is given by

$$\alpha(\theta, \theta') = P(\text{accept} \mid \Theta = \theta, \Theta' = \theta') = \min(1, r(\theta, \theta')).$$

Theorem 9. *The Metropolis–Hastings sampler is reversible with respect to π, and hence has π as its invariant density.*

Proof. To prove reversibility, we must prove that

$$P_\pi(\Theta \in A, \Phi \in B) = P_\pi(\Theta \in B, \Phi \in A) \tag{11.9}$$

for all sets A and B in the state space. Here the subscript π means that the current state Θ is distributed according to the density π.

Now the left-hand side (LHS) of the claim (11.9) is

$$\begin{aligned} P_\pi(\Theta \in A, \Phi \in B) &= P_\pi(\Theta \in A, \Phi \in B, \text{accept}) + P_\pi(\Theta \in A, \Phi \in B, \text{reject}) \\ &= P_\pi(\Theta \in A, \Theta' \in B, \text{accept}) + P_\pi(\Theta \in A \cap B, \text{reject}) \end{aligned}$$

Similarly, the right-hand side (RHS) of the claim (11.9) is

$$P_\pi(\Theta \in B, \Phi \in A) = P_\pi(\Theta \in B, \Theta' \in A, \text{accept}) + P_\pi(\Theta \in B \cap A, \text{reject})$$

The contributions from rejection are equal on the LHS and on the RHS, and we need only show that the contributions from acceptance are also equal.

On the LHS, the contribution from acceptance is

$$\begin{aligned} P_\pi(\Theta \in A, \Theta' \in B, \text{accept}) &= \int \mathrm{d}\theta \, 1_A(\theta) \, \pi(\theta) \int \mathrm{d}\theta' \, 1_B(\theta') \, q(\theta' \mid \theta) \, \alpha(\theta, \theta') \\ &= \iint_{(\theta,\theta') \in A \times B} \pi(\theta) \, q(\theta' \mid \theta) \, \alpha(\theta, \theta') \, \mathrm{d}\theta \, \mathrm{d}\theta'. \end{aligned}$$

Similarly, on the RHS, the contribution from acceptance is

$$\begin{aligned} P_\pi(\Theta \in B, \Theta' \in A, \text{accept}) &= \iint_{(\theta,\theta') \in B \times A} \pi(\theta) \, q(\theta' \mid \theta) \, \alpha(\theta, \theta') \mathrm{d}\theta \, \mathrm{d}\theta' \\ &= \iint_{(\theta,\theta') \in A \times B} \pi(\theta') \, q(\theta \mid \theta') \, \alpha(\theta', \theta) \, \mathrm{d}\theta \, \mathrm{d}\theta', \end{aligned}$$

where in the last formula we just interchanged the names of the integration variables. Since the two integration sets are the same, and the equality has to hold for every integration set $A \times B$, the integrands must be proved to be the same, i.e., the claim (11.9) is true if and only if

$$\pi(\theta) \, q(\theta' \mid \theta) \, \alpha(\theta, \theta') = \pi(\theta') \, q(\theta \mid \theta') \, \alpha(\theta', \theta) \qquad \forall \theta, \theta', \tag{11.10}$$

(almost everywhere). However, our choice (11.8) for $r(\theta', \theta)$ implies (11.10), since its LHS is

$$\begin{aligned} \pi(\theta) \, q(\theta' \mid \theta) \, \alpha(\theta, \theta') &= \pi(\theta) \, q(\theta' \mid \theta) \, \min(1, r(\theta, \theta')) \\ &= \min\left(\pi(\theta) \, q(\theta' \mid \theta), \pi(\theta) \, q(\theta' \mid \theta) \, \frac{\pi(\theta') \, q(\theta \mid \theta')}{\pi(\theta) \, q(\theta' \mid \theta)} \right) \\ &= \min(\pi(\theta) \, q(\theta' \mid \theta), \pi(\theta') \, q(\theta \mid \theta')), \end{aligned}$$

and its RHS is

$$\begin{aligned} \pi(\theta') \; q(\theta \mid \theta') \, \alpha(\theta', \theta) &= \pi(\theta') \, q(\theta \mid \theta') \, \min(1, r(\theta', \theta)) \\ &= \min\left(\pi(\theta') \, q(\theta \mid \theta'), \pi(\theta') \, q(\theta \mid \theta') \frac{\pi(\theta) \, q(\theta' \mid \theta)}{\pi(\theta') \, q(\theta \mid \theta')} \right). \end{aligned}$$

and therefore the two integrands are the same. □

Recall from the proof, that it is sufficient to show the reversibility of the acceptance part of the transition kernel by establishing (11.10), where $\alpha(\theta, \theta') = \min(1, r(\theta, \theta'))$. The formula (11.8) is not the only choice for r which works. E.g., Barker's formula

$$r(\theta, \theta') = \frac{\pi(\theta') q(\theta \mid \theta')}{\pi(\theta') q(\theta \mid \theta') + \pi(\theta) q(\theta' \mid \theta)}$$

(which was proposed by Barker in 1965) would also imply eq. (11.10). Indeed, Hastings considered Barker's formula and many other related formulas for $\alpha(\theta, \theta')$, which all guarantee (11.10). Later, Hasting's student Peskun showed that the acceptance probability $\alpha(\theta, \theta')$ implied by (11.8) is, in a certain sense,

the best possible [9]. Later, Tierney [13] extended Peskun's optimality argument from the discrete state space to the general state space.

If we use a Metropolis–Hastings update to update the jth component of θ only, then the corresponding kernel is reversible with respect to π and hence has π as its invariant density. This follows from our proof, when we treat the other components θ_{-j} as constants. We can then combine the jth component Metropolis–Hastings updates using a systematic scan or a random-scan strategy, and the resulting algorithm still has π as its invariant density. The random scan algorithm is still reversible with respect to π, but the systematic scan algorithm is usually not reversible.

11.7 State-dependent mixing of proposal distributions

As in Sec. 7.4.6 we calculate the proposal θ' as follows, when the current state is θ. We draw the proposal from a proposal density, which is selected randomly from a list of alternatives, and the selection probabilities are allowed depend on the current state.

- Draw j from the pmf $\beta(\cdot \mid \theta), j = 1, \ldots, K$.
- Draw θ' from the density $q(\theta' \mid \theta, j)$ which corresponds to the selected j.
- Accept the proposed value as the new state, if $U < r$, where $U \sim \mathrm{Uni}(0,1)$, and
$$r = \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}. \tag{11.11}$$
Otherwise the chain stays at θ.

We now outline the proof why this yields a Markov chain which is reversible with respect to the target density $\pi(\theta)$.

As in ordinary Metropolis–Hastings, we only need to show reversibility when that the proposed value is accepted. That is, we need to show that

$$P_\pi(\Theta \in A, \Theta' \in B, \text{accept}) = P_\pi(\Theta \in B, \Theta' \in A, \text{accept}), \tag{11.12}$$

where the subscript indicates that the density of the current state Θ is assumed to be π.

Let

$$\begin{aligned}\alpha_j(\theta, \theta') &= P(\text{accept} \mid \Theta = \theta, \Theta' = \theta', \text{component } j \text{ was selected}) \\ &= \min\left(1, \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}\right).\end{aligned}$$

The LHS of the condition (11.12) is

$$\begin{aligned}&\int \mathrm{d}\theta\, 1_A(\theta)\,\pi(\theta) \sum_{j=1}^{K} \beta(j \mid \theta) \int \mathrm{d}\theta'\, q(\theta' \mid \theta, j)\,\alpha_j(\theta, \theta')\,1_B(\theta') \\ &\qquad = \sum_j \iint 1_A(\theta)\,1_B(\theta')\,\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)\,\alpha_j(\theta, \theta')\,\mathrm{d}\theta\,\mathrm{d}\theta'\end{aligned}$$

Similarly, the RHS of the condition (11.12) is

$$\sum_j \iint 1_B(\theta)\, 1_A(\theta')\, \pi(\theta)\, \beta(j \mid \theta)\, q(\theta' \mid \theta, j)\, \alpha_j(\theta, \theta')\, \mathrm{d}\theta\, \mathrm{d}\theta'$$
$$= \sum_j \iint 1_A(\theta)\, 1_B(\theta')\, \pi(\theta')\, \beta(j \mid \theta')\, q(\theta \mid \theta', j)\, \alpha_j(\theta', \theta)\, \mathrm{d}\theta\, \mathrm{d}\theta'$$

The equality of LHS and RHS follows from the fact that the integration sets and the integrands are the same for each j, thanks to the formula (11.11) for the test ratio r.

11.8 Reversibility of RJMCMC

Recall that the reversible jump MCMC method (RJMCMC) allows transitions between parameter spaces of different dimensions. Green derived the RJMCMC algorithm starting from the requirement that the Markov chain should be reversible [4].

We consider reversibility proof for a simple case of the RJMCMC algorithm, where we have two alternative Bayesian models for the same data y. The setting is the same as in Sec. 10.7. The first model is indicated by $M = 1$ and the second model by $M = 2$. The two models have separate parameter vectors θ_1 and θ_2 which we assume to have different dimensionalities d_1 and d_2. Their values are in respective parameter spaces S_1 and S_2. The prior distributions within the two models are

$$p(\theta_1 \mid M = 1), \qquad p(\theta_2 \mid M = 2),$$

and the likelihoods are

$$p(y \mid M = 1, \theta_1), \qquad p(y \mid M = 2, \theta_2).$$

The RJMCMC algorithm constructs a Markov chain, whose state space is the sum space

$$S = (\{1\} \times S_1) \cup (\{2\} \times S_2).$$

Any point in S is of the form (m, θ_m), where m is either 1 or 2, and $\theta_m \in S_m$. The target distribution $\pi(m, \theta_m)$ of the chain is the posterior distribution

$$\pi(m, \theta_m) = p(M = m, \theta_m \mid y), \qquad m = 1, 2, \quad \theta_m \in S_m. \tag{11.13}$$

We suppose that the parameters θ_1 and θ_2 both have continuous distributions and that $d_1 < d_2$.

When the current state of the chain is (m, θ_m), then the algorithm chooses with probability $\beta(m \mid m)$ to attempt to move within the model m or with complementary probability $\beta(k \mid m)$ to attempt to move from the current model m to the other model $k \neq m$.

Recall that in RJMCMC, the moves $1 \to 2$ and $2 \to 1$ must be related in a certain way. Suppose that the move $1 \to 2$ is effected by the following steps, when the current state is $(1, \theta_1)$.

- Draw u_1 from density $g(\cdot \mid \theta_1)$.

- Calculate $\theta_2 = T_{1\to 2}(\theta_1, u_1)$.

We suppose that the function $T_{1\to 2}$ defines a diffeomorphic correspondence between θ_2 and (θ_1, u_1). The density of the noise $g(u_1 \mid \theta_1)$ is a density on the space of dimension $d_2 - d_1$. The test ratio is calculated as

$$r = \frac{\pi(2,\theta_2)}{\pi(1,\theta_1)} \frac{\beta(1 \mid 2)}{\beta(2 \mid 1)} \frac{1}{g(u_1 \mid \theta_1)} \left| \frac{\partial \theta_2}{\partial(\theta_1, u_1)} \right|, \qquad \text{(move } 1 \to 2\text{)}. \tag{11.14}$$

Our choice for the move $1 \to 2$ implies that the move $2 \to 1$ has to deterministic and has to be calculated by applying the inverse transformation $T_{1\to 2}^{-1} = T_{2\to 1}$ to θ_2, when the current state is $(2, \theta_2)$, i.e.,

$$(\theta_1, u_1) = T_{2\to 1}(\theta_2).$$

The value u_1 is also calculated from this requirement, and it is used when we evaluate the test ratio, which is given by

$$r = \frac{\pi(1,\theta_1)}{\pi(2,\theta_2)} \frac{\beta(2 \mid 1)}{\beta(1 \mid 2)} \frac{g(u_1 \mid \theta_1)}{1} \left| \frac{\partial(\theta_1, u_1)}{\partial \theta_2} \right|, \qquad \text{(move } 2 \to 1\text{)}. \tag{11.15}$$

The moves within the models are ordinary Metropolis–Hastings moves from some suitable proposal distributions and for them the test ratio is the ordinary M–H ratio.

To show that RJMCMC is reversible with respect to the target distribution, we should prove that

$$\begin{aligned} P_\pi(M^{(0)} = m, &\Theta^{(0)} \in A, M^{(1)} = k, \Theta^{(1)} \in B) \\ &= P_\pi(M^{(0)} = k, \Theta^{(0)} \in B, M^{(1)} = m, \Theta^{(2)} \in A) \end{aligned} \tag{11.16}$$

for all $m, k \in \{1, 2\}$ and all sets $A \in C_m$ and $B \in C_k$. Here $(M^{(i)}, \Theta^{(i)})$ is the state of the chain at iteration i, and the initial distribution is the target distribution π.

We consider the case $m = 1$ and $k = 2$, and leave the other cases for the reader to check. Let $A \in C_1$ and $B \in C_2$ be arbitrary sets. If the event on the LHS of (11.16) has taken place, then the move $1 \to 2$ has been selected and θ_2 has been proposed and accepted. Therefore the LHS is

$$\int \mathrm{d}\theta_1 \, 1_A(\theta_1) \, \pi(1,\theta_1) \, \beta(2 \mid 1) \int \mathrm{d}u_1 \, g(u_1 \mid \theta_1) \, \min(1, r_{1\to 2}(\theta_1, u_1, \theta_2)) \, 1_B(\theta_2),$$

where $r_{1\to 2}(\theta_1, u_1, \theta_2)$ is the expression (11.14), and θ_2 is short for $T(\theta_1, u_1)$. On the other hand, the RHS is given by

$$\int \mathrm{d}\theta_2 \, 1_B(\theta_2) \, \pi(2,\theta_2) \, \beta(1 \mid 2) \, \min(1, r_{2\to 1}(\theta_2, \theta_1, u_1)) \, 1_A(\theta_1)$$

where $r_{2\to 1}(\theta_2, \theta_1, u_1)$ is the expression (11.15), and the pair (θ_1, u_1) is short for $T_{2\to 1}(\theta_2) = T_{1\to 2}^{-1}(\theta_2)$. Make the change of variables from θ_2 to $(\theta_1, u_1) = T_{1\to 2}^{-1}(\theta_2)$. This changes the RHS to

$$\int \mathrm{d}\theta_1 \int \mathrm{d}u_1 \, 1_A(\theta_1) \, 1_B(\theta_2) \, \pi(2,\theta_2) \, \beta(1 \mid 2) \, \min(1, r_{2\to 1}(\theta_2, \theta_1, u_1)) \left| \frac{\partial \theta_2}{\partial(\theta_1, u_1)} \right|$$

where now θ_2 is short for $T(\theta_1, u_1)$. Taking into account the formulas for the test ratios and remembering that

$$\frac{\partial(\theta_1, u_1)}{\partial \theta_2} \frac{\partial \theta_2}{\partial(\theta_1, u_1)} = 1$$

(since the mappings are inverses of one another) it is routine matter to check that the integrands are the same, and therefore reversibility has been checked for the case $(m, k) = (1, 2)$.

11.9 Irreducibility

A Markov chain which has the target distribution as its invariant distribution may still be useless. For example, consider the trivial Markov chain which stays for ever at the same state where it starts. For this chain, any probability distribution on the state space is an invariant distribution. At the same time, this kernel is clearly useless for the purpose of generating samples from the target distribution. In order to be useful, a Markov chain should visit all parts of the state space. Irreducible chains have that desirable property. A Markov chain which is not irreducible is called reducible.

If the Markov chain has π as it invariant density, then it is called **irreducible**, if for any $\theta^{(0)} \in S$ and for any A such that $\int_A \pi(\theta)\, \mathrm{d}\theta > 0$ there exists an integer m such that

$$P(\Theta^{(m)} \in A \mid \Theta^{(0)} = \theta^{(0)}) > 0.$$

In other words, starting from any initial value, an irreducible chain can eventually reach any subset of the state space (which is relevant for π) with positive probability.

The Metropolis–Hastings sampler (which treats θ as a single block) is irreducible, e.g., if the proposal density is everywhere positive, i.e., if

$$q(\theta' \mid \theta) > 0 \qquad \forall \theta, \theta' \in S.$$

Then every set A which has positive probability under π can be reached with positive probability in one step starting from any θ. However, the positivity of the proposal density is not necessary for the irreducibility of the Metropolis–Hastings chain. It is sufficient that the proposal density allows the chain to visit any region of the space after a finite number of steps.

The jth component Gibbs sampler is, of course, reducible, since it can not change any other components than θ_j. By combining the component updates with a systematic or a random scan strategy, one usually obtains an irreducible chain. The same considerations apply to the Metropolis–Hastings sampler which uses componentwise transitions. However, irreducibility of the Gibbs sampler is not automatic, as the following example shows.

Example 11.1. Let $0 < p < 1$ and consider the density

$$\pi(\theta_1, \theta_2) = p\, 1_{[0,1]\times[0,1]}(\theta_1, \theta_2) + (1 - p)\, 1_{[2,3]\times[2,3]}(\theta_1, \theta_2).$$

The full conditional of θ_1 is the uniform distribution on $[0, 1]$, if $0 < \theta_2 < 1$ and the uniform distribution on $[2, 3]$, if $2 < \theta_2 < 3$. The full conditional of

θ_2 is similar. If we start the simulation using an initial value inside the square $[0,1] \times [0,1]$, then all the subsequent values of the Gibbs sampler will be inside the same square, and the square $[2,3] \times [2,3]$ will never be visited. On the other hand, if we start the simulation using an initial value inside the other square $[2,3] \times [2,3]$, then all the subsequent values of the Gibbs sampler will be inside the same square, and the square $[0,1] \times [0,1]$ will never be visited.

For this target distribution the Gibbs sample is reducible. This example has also the interesting feature that the two full conditional distributions do not determine the joint distribution, since all the joint distributions corresponding to the different $0 < p < 1$ have the same full conditional distributions. $\triangle$

The behavior of the previous example is be ruled out, if the target distribution satisfies what is known as the **positivity condition**. It requires that $\pi(\theta)$ is strictly positive for every θ for which each of the marginal densities of the target distribution $\pi(\theta_j)$ is positive. Thus the support of π has to be the Cartesian product of the supports of the marginal densities. The previous example clearly does not satisfy the positivity condition, since the Cartesian product of the supports of the marginal densities is

$$([0,1] \cup [2,3]) \times ([0,1] \cup [2,3]),$$

but $\pi(\theta) = 0$ for any $\theta \in [0,1] \times [2,3]$ or any $\theta \in [2,3] \times [0,1]$.

The positivity condition ensures irreducibility of the Gibbs sampler, since it allows transitions between any two values in a single cycle. The famous Hammersley–Clifford theorem shows that if the positivity condition is satisfied, then the full conditional distributions determine the joint distribution uniquely.

11.10 Ergodicity

A Markov chain which has an invariant density π is ergodic, if it is irreducible, aperiodic and Harris recurrent. Then the invariant density is unique. Of these conditions, irreducibility has already been discussed.

A Markov chain with a stationary density π is **periodic** if there exist $d \geq 2$ disjoint subsets $A_1, \ldots, A_d \subset S$ such that

$$\int_{A_1} \pi(\theta)\, \mathrm{d}\theta > 0,$$

and starting from A_1 the chain always cycles through the sets $A_1, A_2, \ldots A_d$. I.e., the chain with transition kernel K is periodic with period d, if for the sets A_i

$$K(\theta, A_{i+1}) = 1, \quad \forall \theta \in A_i, \qquad i = 1, \ldots, d-1$$

and

$$K(\theta, A_1) = 1, \quad \forall \theta \in A_d.$$

If the chain is not periodic then it is **aperiodic**. Aperiodicity holds virtually for any Metropolis–Hastings sampler or Gibbs sampler.

The chain is **Harris recurrent**, if for all A with $\int_A \pi(\theta)\, \mathrm{d}\theta > 0$, the chain will visit A infinitely often with probability one, when the chain starts from any initial state $\theta \in S$. For MCMC algorithms, irreducibility usually implies Harris

recurrence, so this property is usually satisfied, although generally irreducibility is a much weaker condition than Harris recurrence.

If the chain is ergodic in the above sense, then starting from any initial value $\Theta^{(0)} = \theta$, the distribution of $\Theta^{(n)}$ converges (in the sense of total variation distance) to the (unique) invariant distribution as n grows without limit.

Under ergodicity, the **strong law of large numbers** holds. Namely, for any real-valued function h, which is absolutely integrable in the sense that

$$\int |h(\theta)|\, \pi(\theta)\, \mathrm{d}\theta < \infty,$$

the empirical means of the RVs $h(\Theta^{(t)})$,

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{t=1}^{n} h(\Theta^{(t)}), \tag{11.17}$$

converge to the corresponding expectation

$$\pi(h) = \int h(\theta)\, \pi(\theta)\, \mathrm{d}\theta \tag{11.18}$$

with probability one, i.e.,

$$\lim_{n \to \infty} \hat{\pi}_n(h) = \pi(h), \tag{11.19}$$

and this holds for any initial distribution for $\Theta^{(0)}$.

11.11 Central limit theorem for Markov chains

We continue to use the notation (11.17) and (11.18). While the central limit theorem (CLT) does not hold for all Markov chains, it does hold for many chains generated by MCMC algorithms. Under regularity conditions on the Markov chain $\Theta^{(i)}$ and integrability conditions for the function h, the CLT then holds for the RVs $h(\Theta^{(i)})$ in the form

$$\sqrt{n}(\hat{\pi}_n(h) - \pi(h)) \xrightarrow{\mathrm{d}} N(0, \sigma_h^2), \qquad \text{as } n \to \infty. \tag{11.20}$$

As a function of the sample size n, the rate of convergence in the Markov chain CLT is the same as in the CLT for i.i.d. random variables. The required conditions on the Markov chain are easiest to state when the chain is reversible with respect to π, and this is why theoreticians recommend that one should favor reversible MCMC algorithms over non-reversible ones. However, these conditions require more advanced notions of ergodicity such as geometric ergodicity, which we bypass. See, e.g., Robert and Casella [10] or Roberts [11] for discussions of the regularity conditions for the CLT.

However, the variance σ_h^2 of the limit distribution is more difficult to estimate than in the i.i.d. setting, since in the Markov chain CLT it is given by the infinite sum

$$\sigma_h^2 = \mathrm{var}_\pi\, h(\Theta^{(0)}) + 2 \sum_{t=1}^{\infty} \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})). \tag{11.21}$$

Here the subscript π means that the covariances are calculated assuming that $\Theta^{(0)} \sim \pi$. Contrast this with the case of i.i.d. sampling from π, where the

variance of the limit distribution would be $\mathrm{var}_\pi\, h(\Theta^{(0)})$. If the chain is extended also for negative times, then this sum can be presented in the doubly-infinite form

$$\sigma_h^2 = \sum_{t=-\infty}^{\infty} \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})),$$

since the autocovariances at lags $-t$ and t are then equal.

One interpretation of the results (11.20) and (11.21) is that we can measure the loss in efficiency due to the use of the Markov chain instead of i.i.d. sampling by defining the parameter

$$\tau_h = \frac{\sigma_h^2}{\mathrm{var}_\pi\, h(\Theta^{(0)})} = 1 + 2\sum_{t=1}^{\infty} \mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})),$$

which is called the **integrated autocorrelation time** for estimating $\pi(h)$ using the Markov chain under consideration (see e.g. [11]). Here $\mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)}))$ is the autocorrelation at lag t for the sequence $\big(h(\Theta^{(t)})\big)$, when the chain is started from the invariant distribution π. We can also define the **effective sample size** (for estimating $\pi(h)$ using the Markov chain under consideration) as

$$n_{\mathrm{eff}}(h) = \frac{n}{\tau_h}$$

This is the sample size of an equivalent i.i.d. sample for estimating $\pi(h)$, when the Markov chain is run for n iterations.

Estimating the asymptotic variance can also be viewed as the problem of estimating the spectral density at frequency zero either for the autocovariance sequence or for the autocorrelation sequence. To simplify the notation, fix the function h and denote the autocovariance sequence of $(h(\Theta^{(t)})$ for the stationary chain by (R_t) and the autocorrelation sequence by (ρ_t),

$$R_t = \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})), \quad \rho_t = \mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})), \qquad t = 0, 1, 2, \ldots$$

Further, let us extend these sequences to negative lags by agreeing that

$$R_{-t} = R_t, \quad \rho_{-t} = \rho_t, \qquad t = 1, 2, \ldots$$

Then the spectral density of the sequence (R_t) at angular frequency w is defined by the Fourier transform

$$g_R(w) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \mathrm{e}^{-itw} R_t, \qquad -\pi < w < \pi,$$

where $i = \sqrt{-1}$. (Warning: there are several related but slightly different definitions of the spectral density in the literature.) The spectral density $g_\rho(w)$ of the sequence (ρ_t) is defined similarly. Using these definitions,

$$\sigma_h^2 = 2\pi\, g_R(0), \qquad \tau_h = 2\pi\, g_\rho(0)$$

There are specialized methods available for the spectral density estimation problem, and these can be applied to estimating the asymptotic variance σ_h^2 or the integrated autocorrelation time τ_h.

All the usual methods for estimating Monte Carlo standard errors in MCMC are ultimately based on the CLT for Markov chains. The methods differ in how one estimates σ_h^2. Some of the methods are based on estimates for the integrated autocorrelation time or of the spectral density at zero. In the batch means method we have already implicitly formed an estimate for σ_h^2. See [2] for further discussion.

11.12 Literature

See the articles [6, 12, 3] and [1, Ch. 14] for surveys of the Markov chain theory needed in MCMC. See the books by Nummelin [7] or by Meyn and Tweedie [5] for comprehensive presentations of the general state space theory. See also the discussions in the books by Robert and Casella [10] and O'Hagan and Forster [8].

Bibliography

[1] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

[2] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260, 2008.

[3] Charles J. Geyer. Introduction to Markov chain Monte Carlo. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC Press, 2011.

[4] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[5] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009. First ed. published by Springer in 1993.

[6] E. Nummelin. MC's for MCMC'ists. *Intenational Statistical Review*, 70(2):215–240, 2002.

[7] Esa Nummelin. *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge University Press, first paperback edition, 2004. First published 1984.

[8] Anthony O'Hagan and Jonathan Forster. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, second edition, 2004.

[9] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[10] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

[11] Gareth O. Roberts. Linking theory and practice of MCMC. In Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

[12] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[13] Luke Tierney. A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8:1–9, 1998.