Markov Chain Monte Carlo

Stochastic Simulation for Bayesian Inference

2nd edition

By Dani Gamerman and Hedibert F. Lopes

2006

Sections 3.1 and 3.4 Monte Carlo Integration

Pages 81-82 & 95-97

CHAPTER 3

Approximate methods of inference

3.1 Introduction

This chapter presents some of the methods proposed for Bayesian inference when the necessary calculations cannot be performed analytically. Some of these techniques are based on deterministic concepts while others are based on non-iterative simulation in opposition to the methods based on iterative simulation that form the core of this text. Therefore, only an introduction to the subject is presented. A more thorough treatment of the subject with comparisons and illustrations of the different techniques is given by Evans and Swartz (1995). The books by Carlin and Louis (2000), Gelman et al. (2004) and O'Hagan and Forster (2004) also provide nice reviews of the area with the first one also providing a summary of software available.

The main techniques presented in this chapter are normal and Laplace approximations based on asymptotics in Section 3.2, quadrature approximations in Section 3.3, Monte Carlo integration in Section 3.4 and resampling techniques in Section 3.5. The last two sections present solutions based on stochastic simulation. They generally involve sampling from an auxiliary distribution that serves different purposes in the context of each approximation.

The deterministic techniques rely upon approximate normality and asymptotic results in the sense of the sample size growing to infinity. These techniques were mostly developed during the 1980s when the computational explosion that enabled computer-intensive methods to be performed was only starting. As will be seen, the complexity of the techniques increases substantially with the dimension of the parametric space. Similar comments are valid for the simulation techniques presented in this chapter. In particular, finding a suitable auxiliary distribution becomes an extremely difficult task. As a consequence, their application to a complete Bayesian analysis in complex models such as those presented at the last sections of the previous chapter is limited. Hierarchical, dynamic and spatial models have in common highly dimensional parameter spaces that are difficult to approach for complete inference with the techniques presented in this chapter.

The last sections are more in the spirit of the book with the use of stochastic simulation for inference from the posterior distribution. The non-iterative form of the simulation used restricts its application in complex

models with large numbers of parameters. For these cases, the use of iterative techniques based on Markov chains and described in all subsequent chapters of this book.

Before going into the details of the techniques, it is important to recall that most summarization operations are provided by integration of the form

$$I = \int t(\theta)\pi(\theta)d\theta. \tag{3.1}$$

The above expression provides the posterior mean of any transformation $\psi = t(\theta)$. When evaluating the posterior mean of θ, $t(\theta) = \theta$. When evaluating the posterior median c of a scalar θ, $t(\theta) = I(\theta < c)$, $I = 1/2$ and (3.1) is solved for c. Similarly, credibility regions are obtained by solving (3.1) for C with $t(\theta) = I(\theta \in C)$ and $I = 1 - \alpha$. The posterior variance matrix may be obtained by taking $t(\theta) = \theta\theta'$ and previously evaluating the posterior mean. Finally, for $\theta = (\theta_1, \ldots, \theta_d)'$ with components θ_i of any dimension, the marginal density of θ_i is given by (2.7). It can be rewritten as

$$\pi(\theta_i) = \int \pi(\theta_i|\theta_{-i})\pi(\theta_{-i})d\theta_{-i} \tag{3.2}$$

and again an integration over a posterior density is required with $t(\theta_{-i}) = \pi(\theta_i|\theta_{-i})$. As mentioned in Section 2.2, another important integral that regularly appears associated with Bayesian model choice and prediction procedures is the posterior predictive density

$$f(y|x) = \int f(y|\theta)\pi(\theta)d\theta$$

which can be easily rewritten as the integral in (3.1) with $t(\theta) = f(y|\theta)$.

In very broad terms, experience gathered from previous work suggests that deterministic techniques provide good results for low dimensional (say single digit) models. Beyond that, they become very complex to handle and Monte Carlo techniques have to be used. When the dimension of the model becomes increasingly large, then only Markov chain simulation seems to provide an adequate solution. Whenever possible, analytical integration should be performed. This will reduce the dimension of the model where approximate methods are applied. Finally, it is important to mention that there is plenty of room for experimentation with combinations of these techniques.

3.4 Monte Carlo integration

Consider as before the problem of solving Equation (3.1). If a sample $\theta_1, \ldots, \theta_n$ from π is available then a natural estimator for I, commonly called the *simple Monte Carlo* (MC) estimator, is

$$\hat{I}_1 = \frac{1}{n} \sum_{j=1}^{n} t(\theta_j).$$

One important application of this result is the derivation of the marginal density of θ_i given by Equation (3.2). A simple MC estimator of this density is obtained by sampling $\theta_{1,-i}, \ldots, \theta_{n,-i}$ from $\pi(\theta_{-i})$ and setting $t(\theta_{-i}) = \pi(\theta_i|\theta_{-i})$. Quite often, sampling from $\pi(\theta)$ (or $\pi(\theta_{-i})$) is either computationally inefficient or costly. Simple Monte Carlo methods must be extended by the use of draws from auxiliary (importance) densities. More specifically, let $q(\theta)$ be another density for θ with the same support of π. Then

$$I = \int \frac{t(\theta)\pi(\theta)}{q(\theta)} q(\theta) d\theta = E_q \left[\frac{t(\theta)\pi(\theta)}{q(\theta)} \right]$$

where E_f denotes expectation with respect to density f. If a sample $\theta_1, \ldots, \theta_n$

from q is available then

$$\hat{I}_2 = \frac{1}{n}\sum_{j=1}^{n}\frac{t(\theta_j)\pi(\theta_j)}{q(\theta_j)} \tag{3.7}$$

is a another estimator of I. $\hat{I}_1$ is a special case of $\hat{I}_2$ obtained when $q = \pi$. Notice that both $\hat{I}_1$ and $\hat{I}_2$ are method of moments estimators of I. These estimators enjoy good frequentist properties such as:

- they are unbiased estimators since $E_q(\hat{I}_k) = I$, for $k = 1, 2$;
- their variances are in the form $V_q(\hat{I}_k) = \sigma_k^2/n$, for $k = 1, 2$, where $\sigma_1^2 = \int[t^2(\theta)\pi(\theta)]d\theta - I^2$ and $\sigma_2^2 = \int[t^2(\theta)\pi^2(\theta)/q(\theta)]d\theta - I^2$;
- they have central limit theorems stating that

$$\sqrt{n}\frac{\hat{I}_k - I}{\sigma_k} \xrightarrow{d} N(0,1) \text{ as } n \to \infty, \tag{3.8}$$

  for $k = 1, 2$; and
- they are strongly consistent estimators of I in that

$$\hat{I}_k \xrightarrow{a.s.} I \text{ as } n \to \infty, \tag{3.9}$$

  for $k = 1, 2$.

The classical nature of the above results leads to objections by Bayesians (O'Hagan, 1987). These results provide important messages, however, and in practice they are widely used. Strong consistency follows directly from the strong law of large numbers (Feller, 1968). So increasing the size n of the sample from q will lead to a virtually error-free estimation at rate $O(n^{-1/2})$. Unlike asymptotic results, this value of n is under the control of the researcher and can be increased by drawing more values from q. The constant σ^2 depends on q and can also be estimated by the method of moments.

The generating density q is usually called the importance density and sampling from q is called importance sampling. There are no restrictions on q and the simplest choice is the uniform distribution when the support of θ is compact. It can be shown that the optimal choice in terms of minimizing σ^2 and hence the estimation error is to take $q \propto t \times \pi$. Unfortunately, for most cases where (3.1) cannot be evaluated analytically, it will be very difficult to draw samples from π. The above results however suggest that q should be taken as close as possible to π but still available for easy sampling. In any case, the importance density q can be chosen to approximate $t \times \pi$ for each required expectation of $t(\theta)$ or can be chosen to be the same for all integrations of interest. Kloek and van Dijk (1978) recommend the latter with the importance density q chosen to approximate π.

Geweke (1989) provides a formal proof of the central limit theorem. It may be used to assess coverage probabilities by confidence intervals thus

providing error bounds for the estimates unlike the previous estimates proposed. Carlin and Louis (2000) and Evans and Swartz (1995) consider this ability as one of the main strengths of approximations based on Monte Carlo techniques.

A problem that usually arises in Bayesian applications is that π is only known up to a proportionality constant. Posterior expectations are really a problem involving a ratio of two integrals as pointed out in (3.5). The resulting approximation is based on the ratio of two Monte Carlo estimators of integrals. Using the same importance density q as recommended above, the numerator and the denominator are approximated by (3.7) with π replaced by $\pi^* = l \times p$ and in the case of the denominator $t = 1$. The form of the estimator is then

$$\hat{I} = \frac{\sum_{i=1}^{n} t(\theta_i)\pi^*(\theta_i)/q(\theta_i)}{\sum_{i=1}^{n} \pi^*(\theta_i)/q(\theta_i)}$$

where the θ_is are the same on numerator and denominator and are sampled from q. The above estimator is only asymptotically unbiased but is still a strongly consistent estimator of I.

Monte Carlo integration has been connected to Bayesian inference after its introduction in applied Econometrics by Kloek and van Dijk (1978). Medium sized models have been commonly used in this area and their paper showed it is a viable technique. Much effort has been devoted since then to the specification of suitable importance density functions. It is important that it matches π as close as possible and that it blankets π in the tails. Otherwise, the very few points sampled in the tails may have large contributions to $\hat{I}$ and estimates will be unstable. This suggests that normal distributions should be avoided if possible.

In the multivariate setting, natural choices for importance density are the Student's t distributions with low degrees of freedom. These are easy to sample, have thick tails and support over R^d. They may therefore require transformation of some of the components of θ to the real line. Geweke (1989) suggested the use of split-t distributions which are obtained by rescaling each component of θ differently for positive and negative values to accommodate skewness. Oh and Berger (1993) suggested the use of mixtures of t-distributions to accommodate posterior multimodality.

These functions require specification of mean and variance which themselves are obtained by integration. This suggests an iterative scheme where means and variances are evaluated for a given importance function and used to update mean and variance specifications of a new importance function. The process is repeated until the successive values of means and variances do not change. Then, integrations of interest are performed. Adaptive strategies have been suggested by Kloek and van Dijk (1978) and Smith et al. (1987). Oh and Berger (1992) established convergence results of these iterative strategies. Examples in Evans and Swartz (1995) suggested that