Chapter 11

Monte Carlo

This chapter introduces the fundamental concepts of Monte Carlo which is both the second MC in MCMC methods and a set of useful techniques used for numerical integration and the assessment of statistical procedures such as estimation and interval construction. Although used in these varied pursuits, Monte Carlo is, fundamentally, a type of numerical approximation to integrals or expected values; we will see that nearly any integral over the real line may be written in the form of an expected value. Such approximations can be used to evaluate integrals for use in a larger procedure, such as maximum likelihood estimation, or to evaluate integrals that quantify the behavior of a statistical estimator or other quantity, such as bias and mean squared error.

11.1 Evaluating Integrals

11.1.1 Integrals as Expected Values

Consider the integral of a continuous function $g(\cdot)$ over all or part of the real line,

$$I = \int_{\mathcal{A}} g(t)\, dt, \tag{11.1}$$

where $\mathcal{A} \subseteq \mathbb{R}$. Let $f(x)$ be a probability density function with support equal to $\mathcal{A}$. Then,

$$I = \int_{\mathcal{A}} \frac{g(t)}{f(t)}\, f(t)\, dt, \tag{11.2}$$

and we have written the integral I as the expected value of $g(t)/f(t)$ with respect to the density $f(\cdot)$. Thus, any integral of the form (11.1) can be written as an expected value with respect to some distribution.

11.1.2 Monte Carlo Evaluation of Integrals

Since any integral may be written as an expected value, consider evaluating an expected value,

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)\, f(x|\boldsymbol{\theta})\, dx, \tag{11.3}$$

where $f(\cdot|\boldsymbol{\theta})$ is a probability density function with parameter $\boldsymbol{\theta} \in \Theta$. We would like a numerical approximation to this expected value, perhaps because the integral cannot be evaluated analytically. Suppose that a random variable with probability density $f(\cdot|\boldsymbol{\theta})$ has finite first absolute moment evaluated at $g(\cdot)$, that is, if X has pdf $f(x|\boldsymbol{\theta})$ then $E|g(X)| < \infty$. Also suppose that, given a value of the parameter $\boldsymbol{\theta}$, we could simulate independent values from the distribution with density $f(\cdot|\boldsymbol{\theta})$. Let these values be denoted $\{X_m^* : \ m =$

$1, \ldots, M\}$. A Monte Carlo approximation to $E\{g(X)\}$ is

$$E_M\{g(X)\} = \frac{1}{M} \sum_{i=1}^{M} g(X_m^*), \tag{11.4}$$

and by the strong law of large numbers (which is why we assumed finite first absolute moment above)

$$E_M\{g(X)\} \xrightarrow{a.s.} E\{g(X)\} \text{ as } M \to \infty. \tag{11.5}$$

The definition of almost sure convergence indicates that we can make $E_M\{g(X)\}$ arbitrarily close to $E\{g(X)\}$ by increasing M.

11.1.3 Monte Carlo Error

The Monte Carlo approximation (11.4) is in the form of a statistical estimator, that is, a sample mean. The expected value of $E_M\{g(X)\}$ is, in fact, the expected value to be approximated, namely $E\{g(X)\}$. With simulated values X_m^* independent and identically distributed we have, in addition to the law of large numbers, the central limit theorm. Let

$$Z_M = \frac{[E_M\{g(X)\} - E\{g(X)\}]}{(var[E_M\{g(X)\}])^{1/2}},$$

and let $H_M(x)$ denote the distribution function of Z_M. Then

$$\lim_{M \to \infty} H_M(x) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}t^2\right) dt, \tag{11.6}$$

for every $x \in \mathbb{R}$. Based on (11.6) we may assess the precision of the Monte Carlo approximation using the interval

$$E_M\{g(X)\} \pm z_{1-\alpha/2} \left(var\left[E_M\{g(X)\}\right]\right)^{1/2}, \tag{11.7}$$

where z_α is the α quantile of a standard normal distribution. Of course to make use of this in application we must have a consistent estimator of

$var[E_M\{g(X)\}]$. An estimator is typically obtained from the the sample variance,

$$S_g^2 = \frac{1}{M-1} \sum_{m=1}^{M} \left[g(X_m^*) - E_M\{g(X)\} \right]^2 . \tag{11.8}$$

An estimator of $var\,[E_M\{g(X)\}]$ is then, in the usual manner for sample means of iid random variables,

$$\hat{V}_M\{g(X)\} = \frac{1}{M} S_g^2 . \tag{11.9}$$

In contrast to other methods in numerical analysis, numerical approximations produced by Monte Carlo methods typically have non-ignorable error. If we approximate an integral (or expected value) using Gaussian quadrature, for example, the approximation is sufficiently precise so that we can ignore any error in the approximation. If we use a Newton Raphson algorithm to find a numerical approximation to a maximum likelihood estimate, we assume the approximation is sufficiently precise that we can ignore any error in the approximation, and this is usually a reasonable assumption. Numerical approximations arrived at by Monte Carlo do not share this characteristic. A good deal of the literature on Monte Carlo evaluation, then, focuses on techniques for reducing error in approximation (see, e.g., Ripley, 1987; Kalos and Whitlock, 1986). We will not address these techniques in detail but rather focus on evaluating the precision that has been obtained in an approximation, and assessing whether Monte Carlo sample size M needs to be increased.

A simple way to assess the precision of a Monte Carlo approximation is to consider meaningful digits as those in which we have a given level of confidence. Consider, for example, a Monte Carlo approximation $E_M[g(X)] = 5.64$, for which M is large and we have estimated the variance $\hat{V}_M\{g(X)\} = 0.0245^2$. We have confidence in the 4 of our estimate if the interval (11.7) is contained in $[5.635, 5.645)$. A 95% interval here is $(5.592, 5.688)$, so we cannot claim

95% confidence in the digit 4 of the approximation 5.64. We assume that digit could be anything between a 3 and a 5. We then have confidence in the 6 of our estimate if an interval computed as $5.6 \pm 1.96(0.0245)$ is contained in $[5.55, 5.65)$. The interval is $(5.552, 5.648) \subset [5.5, 5.65)$ so we can claim 95% confidence in the digit 6. We rarely undertake such calculations because $\hat{V}_M\{g(x)\}$ is itself estimated, but reporting the Monte Carlo variance allows any interested party to make their own assessment of the quality of the Monte Carlo approximation being reported.

11.1.4 Importance Sampling

It is not uncommon that we wish to approximate an integral that is already in the form of an expected value, but for which sampling from the appropriate distribution is difficult or impossible. This may occur, for example, if the distribution corresponds to a complex model structure, or is a distribution not included in a standard set of random number generating functions in a computer language or software package. Suppose, then, that our objective is to approximate the integral

$$E(X) = \int_{-\infty}^{\infty} x\, f(x|\boldsymbol{\theta})\, dx, \tag{11.10}$$

but it would be difficult to simulate values from $f(x|\boldsymbol{\theta})$. We may apply the same device used previously to express any integral as an expected value. Suppose we are able to identify another distribution with density $h(x|\boldsymbol{\lambda})$ such that the support of h dominates the support of f, meaning that $h(x) > 0$ for every x such that $f(x) > 0$. We attempt to identify an h that is easier to sample from than is f. Then we can rewrite (11.10) as

$$E(X) = \int_{-\infty}^{\infty} x\, \frac{f(x|\boldsymbol{\theta})}{h(x|\boldsymbol{\lambda})}\, h(x|\boldsymbol{\lambda})\, dx. \tag{11.11}$$

The quantity to be approximated has now been written as an expected value with respect to a distribution from which we can easily sample. We would then sample values $\{X_m^* : \ m = 1, \ldots, M\}$ from $h(x|\boldsymbol{\lambda})$ and a Monte Carlo approximation to $E(X)$ would be computed as

$$E_M(X) = \frac{1}{M} \sum_{m=1}^{M} X_m^* \frac{f(X_m^*|\boldsymbol{\theta})}{h(X_m^*|\boldsymbol{\lambda})}. \tag{11.12}$$

In this procedure, called importance sampling, the distribution that is sampled (simulated) from, $h(x|\boldsymbol{\lambda})$ is called the importance distribution or the importance sampling distribution.

Importance sampling was originally used by numerical analysts as a technique for reducing the variance of Monte Carlo approximations. Today, statisticians often use it in the context of this subsection, when the integral contains a density that is difficult to sample from at all. It also bears mention that importance sampling can be used to allow Monte Carlo approximation to integrals that contain unbounded densities such as a beta distribution with parameters $\alpha = 1/2$ and $\beta = 1/2$. A straight Monte Carlo approximation to an expectation with respect to an unbounded density can have variance ∞, but importance sampling can be used to produce an approximation with finite variance.

11.2 Reporting Results of Monte Carlo Studies

A traditional use of Monte Carlo methods is to conduct investigations of the behavior of statistical estimators, tests, or other procedures. A simulation study may, for example, evaluate the bias and variance or mean squared error of several competing estimators for a class of problems. A simulation study may focus on evaluating the size (type I error rate) and power of a test procedure. Or, a simulation study may evaluate the sample size necessary for an asymptotic result to provide a good approximation to the sampling distribution of an estimator. In these and similar applications of Monte Carlo we need to avoid misinterpretation of results and confusion caused by semantics. A fundamental difficulty is determining what is meant by the phrase "Monte Carlo variance". Is a Monte Carlo variance the sample variance of Monte Carlo values as in (11.8)? Is a Monte Carlo variance the variance of a Monte Carlo approximation as in (11.9)? Or is a Monte Carlo variance a Monte Carlo approximation to an expected variance? A good deal of the confusion caused by imprecise semantics can be avoided by using the phrases (1) Monte Carlo approximation rather than a Monte Carlo estimate, and (2) variance of the Monte Carlo approximation rather than Monte Carlo variance.

Although the quantities of interest in a Monte Carlo study can vary considerably, as a general rule it is useful to report both the Monte Carlo approximation and an interval approximation. For a straightforward Monte Carlo study these are given by expressions (11.4) and (11.7), respectively. Alternatively, and particularly when the expected values of a number of quantities are being approximated, we might compute and report intervals for a number of sample sizes with one of the basic quantities under investigation and, based on those results, pick a fixed Monte Carlo sample size for the study.

Example 11.1

As an example of a simple Monte Carlo study, consider an investigation into the relative behaviors of exact and approximate normal theory interval estimators of a mean. We will make use of two models in the simulation, formulated for two groups of independent random variables, $\{Y_{1,i} : \ i = 1, \ldots, n_1\}$ and $\{Y_{2,i} : \ i = 1, \ldots, n_2\}$.

$$\begin{aligned} M1 &: Y_{1,i} \sim iid \ \mathrm{N}(\mu_1, \sigma_1^2) \quad Y_{2,i} \sim iid \ \mathrm{N}(\mu_2, \sigma_2^2) \\ M2 &: Y_{1,i} \sim iid \ \mathrm{N}(\mu_1, \sigma^2) \quad Y_{2,i} \sim iid \ \mathrm{N}(\mu_2, \sigma^2) \end{aligned} \tag{11.13}$$

Model $M1$ specifies normal distributions with different means and variances, while model $M2$ specifies normal distributions with different means but equal variance. For each data set simulated from one of these models, two interval estimates were computed.

1. An approximate interval for the difference in means was computed as

$$\bar{Y}_1 - \bar{Y}_2 \pm z_{1-\alpha/2} \left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right]^{1/2}, \tag{11.14}$$

   where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of a standard normal distribution (i.e., $N(0, 1)$). This interval can be justified in a number of ways, perhaps the simplest being maximum likelihood for regular problems. Under either model $M1$ or model $M2$, $\bar{Y}_1$ and $\bar{Y}_2$ are maximum likelihood estimators and have independent normal distributions. By invariance, the difference $\bar{Y}_1 - \bar{Y}_2$ is then also a maximum likelihood estimator and has a normal distribution for any sample sizes n_1 and n_2. The quantity

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right]^{1/2}}$$

   has an asymptotically normal distribution, from which the approximate interval (11.14) follows.

2. An exact interval for the difference in means was computed as

$$\bar{Y}_1 - \bar{Y}_2 \pm t \left[\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}\right]^{1/2}, \tag{11.15}$$

   where t is the $(1-\alpha/2)$ quantile of a $t-$distribution with $n_1 + n_2 - 2$ degrees of freedom and S_p^2 is the usual pooled sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

   This interval is justified using exact normal theory under model $M2$.

The behaviors of these estimators we will examine are:

1. How well the actual coverage of interval estimates matches the specified nominal level.

2. The average width of interval estimates under the different assumptions used in their computation. Smaller widths indicate more precise intervals, **given that they have the correct coverage**.

To select a Monte Carlo sample size for this example, approximations and interval approximations for coverage of approximate intervals for the difference in means (11.14) were computed as Monte Carlo sample size increased from $M = 500$ to $M = 100000$. Data were simulated from model $M1$ with $\mu_1 = 15$, $\mu_2 = 20$, $\sigma_1^2 = 4$, $\sigma_2^2 = 64$, and $n_1 = n_2 = 50$. The results are presented in Table 11.1 The values in Table 11.1 point out several aspects of the effect of Monte Carlo sample size that are typical in simulation studies. First, note that the precision of Monte Carlo approximations to coverage rate increases fairly rapidly as M increases to about 2000 or 3000, but then the improvement in precision slows considerably with additional increases in sample size. Despite that fact, however, it is not until quite large sample sizes that we begin to see that our approximate interval has coverage just slightly less than the nominal rate of 95%, presumably because despite the fact that data were simulated from normal distributions, the result on which the interval (11.14) is based is asymptotic and the statistical sample size used was $n_1 = n_2 = 50$. Whether this is of practical importance or not might well depend on the problem. For many purposes, a Monte Carlo sample size of $M = 5000$ appears more than adequate to get a handle on the behavior of our interval estimators.

M	MC Approx.	MC Interval	Width
500	0.936	(0.914, 0.957)	0.043
1000	0.942	(0.928, 0.956)	0.029
2000	0.948	(0.938, 0.957)	0.020
3000	0.952	(0.944, 0.959)	0.015
4000	0.949	(0.942, 0.956)	0.014
5000	0.949	(0.943, 0.955)	0.012
6000	0.948	(0.941, 0.952)	0.011
7000	0.951	(0.946, 0.957)	0.010
8000	0.949	(0.944, 0.954)	0.010
9000	0.948	(0.944, 0.953)	0.009
10000	0.947	(0.943, 0.951)	0.008
20000	0.948	(0.945, 0.951)	0.006
50000	0.947	(0.945, 0.949)	0.004
100000	0.947	(0.945,0.948)	0.003
200000	0.946	(0.945, 0.947)	0.002

Table 11.1: Monte Carlo approximations to coverage of 95% approximate intervals for the difference in two normal means.

Results for data sets simulated from model $M1$ (unequal variance) are presented in Table 11.2. Examine the first three lines of this table. The only difference in simulation structure is that the discrepancy in variances was increased, until variances for the two groups were quite dissimilar. The approximate interval maintained fairly consistent coverage at about $0.940 - 0.947$. Despite the differences in variance, the exact estimator maintained a coverage of $0.945 - 0.950$, even closer to the nomial level of 0.95 than the approximate interval, which should actually be more appropriate given the unequal variances. Overall, there appears little to choose from in terms of preferring one of these estimators over the other. Now examine the last three rows of Table 11.2. These rows repeat the differences in variances from the first three rows, but now with unequal sample sizes, $n_1 = 50$ but $n_2 = 25$. Here, we see a tiny drop in coverage for the approximate interval, but a large decrease in coverage for the exact interval estimator. Apparently, it is not wise to make use of a pooled estimate of sample variance when both sample sizes and variances differ.

Another Monte Carlo simulation was run using model $M2$ as a data generating mechanism. Results are presented in Table 11.3. In this case, common means and variances were used, with differences between rows being due only to a decrease in sample size. The main point of Table 11.3 is that the sample sizes need to be reduced to quite small values (3 or maybe 5) before the approximate interval suffers a serious degradation in coverage. The benefit of using a pooled sample variance, of course, is that one increases the number of observations used in variance estimation. This benefit is seen, but not until sample sizes are reduced to quite low levels. This suggests that a procedure consisting of a test for equality of variance before assuming equal variance in a comparison of means is without force. If one has sufficient sample size for

						Coverage		Width	
μ_1	μ_2	σ_1^2	σ_2^2	n_1	n_2	Approx.	Exact	Approx.	Exact
15	15	4	64	50	50	0.944	0.947	1.75	1.77
15	15	4	100	50	50	0.941	0.945	1.91	1.94
15	15	4	200	50	50	0.947	0.950	2.21	2.24
15	15	4	64	50	25	0.944	0.888	2.33	1.94
15	15	4	100	50	25	0.944	0.883	2.58	2.09
15	15	4	200	50	25	0.939	0.869	3.03	2.38

Table 11.2: Monte Carlo results for comparison of approximate and exact interval estimates with data generated from model $M1$ using Monte Carlo sample size $M = 5000$.

a test of equality of variance to be meaningful, one has sufficient sample size that it really doesn't matter if one assumes equal variance or not. So how does one justify an assumption of equal variance between or among groups? One cannot use the data to justify this assumption, one must use the design of the study. This essentially requires an experimental setting where all relevant factors are under the control of the investigator and only the active treatment(s) are allowed to differ among groups. Variance is then primarily observational error which for many measuring instruments tend to be symmetric and, as long as the same instrument is used for all treatment groups, should be equal among groups.

One of the most common uses of Monte Carlo studies is to investigate the behavior of one or more statistical point estimators of a parameter $\boldsymbol{\theta}$. For simplicity we will assume that this is a scalar parameter θ and the estimators will be denoted as $\hat{\theta}$ and $\tilde{\theta}$. The following lists Monte Carlo approximations

						Coverage		Width	
μ_1	μ_2	σ_1^2	σ_2^2	n_1	n_2	Approx.	Exact	Approx.	Exact
15	15	16	16	50	50	0.954	0.957	1.56	1.58
15	15	16	16	10	10	0.937	0.950	3.46	3.70
15	15	16	16	5	5	0.920	0.958	4.29	5.68
15	15	16	16	3	3	0.879	0.949	6.03	8.54

Table 11.3: Monte Carlo results for comparison of approximate and exact interval estimates with data generated from model $M2$ using Monte Carlo sample size $M = 5000$.

that are commonly used in this situation.

1. Expected values $E_M(\hat{\theta})$ and $E_M(\tilde{\theta})$ or biases $B_M(\hat{\theta}) = E_M(\hat{\theta}) - \theta$ and $B_M(\tilde{\theta}) = E_M(\tilde{\theta}) - \theta$.

2. Variances $E_M\left\{\hat{\theta} - E_M(\hat{\theta})\right\}^2$ and $E_M\left\{\tilde{\theta} - E_M(\tilde{\theta})\right\}^2$.

3. Mean squared errors $mse_M(\hat{\theta}) = E_M\left\{\hat{\theta} - \theta\right\}^2$ and $mse_M(\tilde{\theta}) = E_M\left\{\tilde{\theta} - \theta\right\}^2$, which as usual are also variance plus squared bias. Root mean squared errors are also often used.

4. Relative biases, $B_M(\hat{\theta})/\theta$ and $B_M(\tilde{\theta})/\theta$.

5. Relative root mean squared errors, $\left[mse_M(\hat{\theta})\right]^{1/2}/\theta$ and $\left[mse_M(\tilde{\theta})\right]^{1/2}/\theta$.

6. Mean (or median) absolute differences, $D_M(\hat{\theta}) = E_M\left(|\hat{\theta} - \theta|\right)$ and $D_M(\tilde{\theta}) = E_M\left(|\tilde{\theta} - \theta|\right)$.

7. Relative mean (or median) absolute differences $D_M(\hat{\theta})/\theta$ and $D_M(\tilde{\theta})/\theta$.

11.3 Monte Carlo in MCMC

As mentioned at the beginning of this chapter, Monte Carlo is the second portion of Markov Chain Monte Carlo procedures. We previously discussed using Markov Chains to approximate posterior distributions using the empirical distributions of simulated values having the appropriate posteriors and in that endeavor the lack of independence among simulated values caused no difficulties. That is, as long as the final empirical distribution reflects the appropriate placement of probability in the posterior distribution, it is not important whether that empirical distribution was constructed using independent values or whether it was built up from all small values first and then larger values later, and so forth. But lack of independence in simulated values vitiates the strong law a large numbers and the central limit theorem in Monte Carlo approximations formed from Markov Chain output. For the Monte Carlo portion of MCMC to be useful in, for example, approximating the posterior expectation with the average of values simulated from a Markov Chain that has the given posterior as its limit distribution we need a version of the law of large numbers that holds for correlated sequences of values. Similarly, to assess the uncertainty in such a Monte Carlo approximation we need a central limit theorem that holds for sequences of dependent random variables. This can become quite a complex topic and deals, in general, with asymptotic behavior called ergodicity. While a full treatment of this topic is beyond the scope of this book, we will attempt to provide a flavor of the issues involved and what is needed to have reasonable confidence in using Monte Carlo summaries of empirical approximations to posterior distributions produced by Markov Chain samplers.

11.3.1 The Effect of Dependence on Monte Carlo Approximations

Consider the random variable X with probability mass or density function f. Assume we are using the simple MC estimator $\hat{G}_M = (1/M)\sum_{i=m}^{M} g(X_m)$ to approximate $G = E[g(X)]$, where $\{X_m : m = 1, \ldots, M\}$ are values simulated from f. If $X_m \sim$ iid f, then

$$\begin{aligned}\text{var}(\hat{G}_M) &= \frac{\sigma^2}{M},\\ \sigma^2 &= E\left([g(X) - E\{g(X)\}]^2\right).\end{aligned}$$

Given σ^2 we could then, for example, determine M such that $\sigma^2/M = \delta$ for some selected value δ. Note here that $\sigma^2 = \text{var}[g(X)]$ which is not necessarily the same asvar(X).

If we do not have independent MC draws X_m, this no longer holds. Suppose that $\rho_s = \text{cor}[g(X_m),\, g(X_{m+s}]$. Then,

$$\begin{aligned}\text{var}(\hat{G}_M) &= \frac{1}{M^2}\sum_{m=1}^{M}\text{var}[g(X_m)] + \frac{2}{M^2}\sum\sum\nolimits_{1\le m<h\le M}\text{cov}[g(X_m), g(X_h)]\\ &= \frac{\sigma^2}{M} + \frac{2\sigma^2}{M^2}\sum_{s=1}^{M-1}(M-s)\rho_s. \qquad (11.16)\end{aligned}$$

If all $\rho_s \ge 0$ and a sufficient number of $\rho_s > 0$ then this variance is greater than σ^2/M, the independence case value and we will need more MC draws to achieve the same variance δ. Now, in neither case do we know σ^2 and we will need to estimate it. A natural estimator is the sample variance,

$$S^2 = \frac{1}{M-1}\sum_{m=1}^{M}\left[g(X_m) - \hat{G}_M\right]^2.$$

Now,

$$\begin{aligned}
E(S^2) &= \frac{1}{M-1}\sum_{m=1}^{M} E\left[g(X_m) - \hat{G}_M\right]^2 \\
&= \frac{1}{M-1}\sum_{m=1}^{M} E\left[\{g(X_m) - G\} - \left\{\hat{G}_M - G\right\}\right]^2 \\
&= \frac{1}{M-1} E\left[\sum_{m=1}^{M} \{g(X_m) - G\}^2 - M\left(\hat{G}_M - G\right)^2\right] \\
&= \frac{1}{M-1}\left[M\sigma^2 - M\text{var}(\hat{G}_M)\right].
\end{aligned}$$

If $\text{var}(\hat{G}_M) = \sigma^2/M$ as for the independence case in which all $\rho_s = 0$ in (11.16), $E(S^2) = \sigma^2$ and S^2/M is unbiased for $\text{var}(\hat{G}_M)$, but if $\text{var}(\hat{G}_M) > \sigma^2/M$, $E(S^2) < \sigma^2$ and S^2/M underestimates even σ^2/M which is itself less than $\text{var}(\hat{G}_M)$.

11.3.2 Improving Estimation of Variances

There are a number of techniques that have been developed to improve the estimation for variance of a Monte Carlo approximation $\text{var}(\hat{G}_M)$. An approach based on time series methods produces estimates of the correlations ρ_s which are then substituted into (11.16). See Flegal and Jones (2011) and Ripley (1987) for more information on this approach. We will consider a different approach based on dividing the sequence of values in a Markov chain into batches or blocks.

Non-Overlapping Blocks

Suppose we divide the values in a realized chain of length M into k non-overlapping batches or blocks, each of size b, so that $kb = M$. The blocks are

$\{B_j : \ j = 1, \ldots, k\}$ where,

$$B_j = \{X_m : \ m = (j-1)b + 1, \ldots, jb\}.$$

For $j = 1, \ldots, k$ let

$$W_j = \frac{1}{b} \sum_{m \in B_j} g(X_m),$$

and

$$\bar{W} = \frac{1}{k} \sum_{j=1}^{k} W_j = \frac{1}{M} \sum_{m=1}^{M} g(X_m).$$

Note that $\bar{W}$ does not depend on k or b other than through $kb = M$. Consider as an estimator of $\mathrm{var}(\hat{G}_M)$ the quantity,

$$\hat{V} = \frac{1}{k(k-1)} \sum_{j=1}^{k} (W_j - \bar{W})^2. \tag{11.17}$$

From (11.16) we have that,

$$\begin{aligned}
\mathrm{var}(\hat{G}_M) = \mathrm{var}(\bar{W}) &= \frac{\sigma^2}{M} + \frac{2\sigma^2}{M^2} \sum_{s=1}^{M-1} (M-s)\rho_s \\
&= \frac{\sigma^2}{M} \left[1 + 2 \sum_{s=1}^{M-1} \left(1 - \frac{s}{M}\right) \rho_s \right] \\
&= \frac{1}{M} \, Q(M, \sigma^2, \boldsymbol{\rho}),
\end{aligned}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{M-1})$. Similarly, for $j = 1, \ldots, k$,

$$\begin{aligned}
\mathrm{var}(W_j) &= \frac{\sigma^2}{b} \left[1 + 2 \sum_{s=(j-1)b+1}^{jb-1} (b-s)\rho_s \right] \\
&= Q(b, \sigma^2, \boldsymbol{\rho}).
\end{aligned}$$

This leads to,

$$E(\hat{V}) = \frac{1}{k(k-1)} E\left[\sum_{j=1}^{k}(W_j - \bar{W})^2\right] \tag{11.18}$$

$$= \frac{1}{k(k-1)} E\left[\sum_{j=1}^{k}(W_j - G)^2 - k(\bar{W} - G)^2\right] \tag{11.19}$$

$$= \frac{1}{k(k-1)} \left[\frac{k}{b} Q(b, \sigma^2, \boldsymbol{\rho}) - \frac{k}{M} Q(M, \sigma^2, \boldsymbol{\rho})\right]. \tag{11.20}$$

Substituting the previous forms for $Q(m, \sigma^2, \boldsymbol{\rho})$ and $Q(M, \sigma^2, \boldsymbol{\rho})$ into (11.18), after a fair bit of algebra it can be shown that,

$$E(\hat{V}) = \text{var}(\hat{G}_M) - \frac{2\sigma}{M-b)} \left[\sum_{s=1}^{b-1} \frac{M-b}{Mb} s\rho_s + \sum_{s=b}^{M-1}\left(1 - \frac{s}{M}\right)\rho_s\right]. \tag{11.21}$$

As M becomes large and if $\rho_s \approx 0$ for $s > b$, $\hat{V}$ is essentially unbiased for $\text{var}(\hat{G}_M)$.

Overlapping Blocks

Blocks may also be formed with overlap, shifting the starting value in each block by one. If a chain is of length M and the block length is b then we will have $k = M - b + 1$ blocks,

$$B_j = \{X_m : m = j, \ldots, b + j - 1\}.$$

As previously, let $W_j = (1/b)\sum_{m \in B_j} g(X_m)$. We still have the Monte Carlo approximation $\hat{G}_M\{g(X)\} = (1/M)\sum_{m=1}^{M} g(X_m)$ computed from the entire chain. Now, let

$$\hat{V} = \frac{Mb}{(M-b)(M-b+1)} \sum_{j=1}^{M-b+1} [W_j - \hat{G}_M]^2. \tag{11.22}$$

It appears that if b is allowed to increase with M (as then b_M), the Markov chain mixes well, and g has a sufficient number of finite moments with respect to the distribution f, then $\hat{V}$ in (11.22) is strongly consistent for $\text{var}(\hat{G}_M)$ as $M \to \infty$ (Brooks, Gelman, Jones and Meng 2011; Flegal and Jones 2010). It is common to take b_M as the greatest integer less than or equal to M^ν with ν often chosen to be $1/4, 1/2$ or $3/4$.

Subsampling

Suppose we have a sequence of values (i.e., a Markov chain) $\{X_m : m = 1, \ldots, M\}$ with a joint probability distribution P_θ indexed by some parameter θ and some estimator of θ, $\hat{\theta}_M = \hat{\theta}(X_1, \ldots, X_M)$. Suppose further that a standardized version of $\hat{\theta}_M$ has a limit distribution,

$$\tau_M(\hat{\theta}_M - \theta) \xrightarrow{d} H. \tag{11.23}$$

We may or may not know H explicitly but we must know it exists. We assume we do know an appropriate sequence of normalizing factors τ_M, but these do not necessarily have to be $M^{1/2}$. In the case that $\theta = \mu = E(X_1)$ for a stationary sequence of values and $\hat{\theta}_M = \bar{X}_M = (1/M)\sum_{m=1}^{M} X_m$, under mild regularity conditions we have the familiar $\sqrt{M}(\bar{X}_M - \mu)$ with a normal limit distribution having some variance other than $\text{var}(X_1)$. If $\theta = E[g(X_1)]$ and $\hat{\theta}_M = (1/M)\sum_{m=1}^{M} g(X_m)$ we have the problem considered in the previous subsection in which approximation of the limiting variance was approached through the use of non-overlapping blocks. These problems fall under our current formulation, but our treatment here is also more general.

Subsampling makes use of overlapping blocks of values, in which the starting value in each block is shifited by one. For asymptotic properties to be developed for what follows, we will need the block size to increase with the chain length and this can be denoted as b_M. We assume that $M \to \infty$, $b_M \to \infty$ and $b_M/M \to 0$. For ease of notation, we will drop the formal dependence of block length on chain length with the understanding that in what follows, $b = b_M$. If a chain is of length M and the block length is b then we will have $k = M - b + 1$ blocks,

$$B_j = \{X_m : m = j, \ldots, b + j - 1\}.$$

The essential concept underlying subsampling is that the limit distribution of the statistic in (11.23) is the same as the limit distribution of the same statistic applied to any one block,

$$\tau_b(\hat{\theta}_{b,j} - \theta),$$

where, for $j = 1, \ldots, k$, $\hat{\theta}_{b,j} = \hat{\theta}(X_j, \ldots, X_{b+j-1})$. The normalizing factor τ_b is the same function of b that τ_M is of M in (11.23). The subsampling approximation to the sampling distribution of the scaled and centered statistic in (11.23) is the empirical distribution function, for any $-\infty < y < \infty$,

$$L_{M,b}(y) = \frac{1}{M - b + 1} \sum_{j=1}^{M-b+1} I[\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M) \leq y], \tag{11.24}$$

where $I(A)$ is the identity function that assumes a value of 1 if A is true, and 0 otherwise.

A $(1 - \alpha)100\%$ confidence interval can be computed directly from (11.24) by determining the $\alpha/2$ and $1 - \alpha/2$ quantiles of $L_{M,b}(y)$,

$$\begin{aligned} q_{M,\alpha/2} &= \min\{y : L_{M,b}(y) \geq \alpha/2\} \\ q_{M,1-\alpha/2} &= \min\{y : L_{M,b}(y) \geq 1 - \alpha/2\}. \end{aligned}$$

The confidence interval then becomes (L, U), where,

$$\begin{aligned} L &= \hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2} \\ U &= \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2}. \end{aligned} \tag{11.25}$$

In practice, let $\nu_1 = \lfloor (M+1)\alpha/2 \rfloor$ and $\nu_2 = \lfloor (M+1)(1-\alpha/2) \rfloor$ where $\lfloor Y \rfloor$ denotes the largest integer less than or equal to Y. Then $q_{M,\alpha/2}$ is approximated by $\tilde{q}_{M,\alpha} = \{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M)\}_{[\nu_1]}$ and $q_{M,1-\alpha/2}$ is approximated by $\tilde{q}_{M,1-\alpha} = \{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M)\}_{[\nu_2]}$, where $X_{[q]}$ denotes the q^{th} ordered value of X. The confidence interval is then given by (11.25) with $\tilde{q}_{M,\alpha}$ in place of $q_{M,alpha}$ and $\tilde{q}_{M,1-\alpha}$ in place of $q_{M,1-\alpha}$.

The centering value $\hat{\theta}_M$ in (11.24) is sometimes replaced by the average of the subsampling estimates, $\bar{\theta}_b = (1/k)\sum_{j=1}^{k} \hat{\theta}_{b,j}$. Whether this is superior or inferior to use of $\hat{\theta}_M$ computed from the entire chain remains a matter of debate, although asymptotically there is no difference.

Another approach is available in the case that (11.23) has the form,

$$\sqrt{M}(\hat{G}_M - E\{g(X_1)\}) \xrightarrow{d} N(0, \sigma_g^2),$$

where $\hat{G}_M = (1/M)\sum_{m=1}^{M} g(X_m)$. In this case we could approximate $\sigma_g^2 = \mathrm{var}(\hat{G}_M)$ as,

$$\hat{\sigma}_g^2 = \frac{1}{M-b+1} \sum_{j=1}^{M-b+1} \sqrt{b}(\hat{G}_{M,j} - \bar{G}_M)^2, \tag{11.26}$$

where $\bar{G}_M = [1/(M+b-1)]\sum_{j=1}^{M+b-1} \hat{G}_{M,j}$. A $(1-\alpha)100\%$ confidence interval for $E\{g(X_1)\}$ may then be computed as (L, U) where

$$L = \hat{G}_M - z_{1-\alpha/2}[\hat{\sigma}_g^2]^{1/2},$$
$$U = \hat{G}_M + z_{1-\alpha/2}[\hat{\sigma}_g^2]^{1/2}.$$

This approach is taken, for example, by Flegal and Jones (2011).

11.4 Case Study: Approximating a Spatial Average

This case study is a little different from many in this book in that it involves an exercise to examine the behavior of several approaches to estimating the probability of the occurrence of some event at spatial locations. We believe the occurrence of the event at locations depends on the value of some covariate, and also that the occurrence of the event at one location influences the probability that the event also occurs at other nearby locations.

11.4.1 Problem Formulation

Consider a finite set of locations on a regular lattice in two-dimensional space, denoted as $\{\boldsymbol{s}_i = (u_i, v_i) : i = 1, \ldots, n\}$ where $u_i \in \{1, \ldots, k\}$ is a horizontal index and $v_i \in \{1, \ldots, k\}$ is a vertical index. The locations $\boldsymbol{s}_i$ can be thought of as the vertices of a $k \times k$ regular grid. At each location $\boldsymbol{s}_i$ there is a response random variable $Y(\boldsymbol{s}_i)$; $i = 1, \ldots, n$ and a spatial covariate x_i; $i = 1, \ldots, n$. The response variables take a value of 1 if the event of interest has occurred at the location and a value of 0 if it has not. We also define what are called neighborhoods for each location and the neighborhood sets $N_i = \{\boldsymbol{s}_j : \boldsymbol{s}_j \text{ is a neighbor of} \boldsymbol{s}_i\}$. Common neighborhood configurations are $4-$nearest and $8-$nearest neighbors. For a $4-$nearest neighborhood structure, $N_i = \{\boldsymbol{s}_j : (u_j = u_i \pm 1, v_j = v_i) \cup (u_j = u_i, v_j = v_i \pm 1)\}$. A model for the response variables is formulated by specifying the full conditional distributions of each $Y(\boldsymbol{s}_i)$ given values at all other locations $\{y(\boldsymbol{s}_j) : j \neq i\}$. Letting $p(\cdot)$ be generic notation for a density or mass function, a Markov assumption is made such that,

$$p(y(\boldsymbol{s}_i)\,|\,\{y(\boldsymbol{s}_j) : j \neq i\}) = p(y(\boldsymbol{s}_i)\,|\,\{y(\boldsymbol{s}_j) \in N_i\}) = p(y(\boldsymbol{s}_i\,|\,\boldsymbol{y}(N_i)). \quad (11.27)$$

That is, the assumption is that the full conditional distribution of $Y(\boldsymbol{s}_i)$ given values at all other locations is the same as the conditional distribution of $Y(\boldsymbol{s}_i)$ given only values at locations in it's neighborhood. In general, given a set of conditional distributions (11.27), one must show at a joint distribution $p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ exists and then use it for estimation and inference. If we choose the conditional distributions to be binary, we might specify a model similar to that of Caragea and Kaiser (2009), for some $-\infty < \beta_0, \beta_1, \eta < \infty$,

$$\begin{aligned} p(y(\boldsymbol{s}_i)\,|\boldsymbol{y}(N_i)) &= \mu_i^{y(\boldsymbol{s}_i)}(1-\mu_i)^{1-y(\boldsymbol{s}_i)};\ y(\boldsymbol{s}_i) = 0, 1 \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \eta \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \kappa_j\} \\ \log\left(\frac{\kappa_i}{1-\kappa_i}\right) &= \beta_0 + \beta_1 x_i. \end{aligned} \quad (11.28)$$

Caragea and Kaiser (2009) demonstrate that for reasonable values of η, κ_i will be close to the marginal expected value of $Y(\boldsymbol{s}_i)$. Exactly what reasonable and close mean here will be left vague for the moment, but see Kaiser, Caragea, and Furukawa (2012).

For a model based on the conditionals (11.28) it can be shown that the joint distribution exists, but that distribution can be identified only up to a constant of proportionality. Specifically, for $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^T$,

$$p(\boldsymbol{y}\,|\,\beta_0, \beta_1, \eta) = \frac{1}{K(\beta_0, \beta_1, \eta)} \exp[Q(\boldsymbol{y}|\beta_0, \beta_1, \eta)]. \quad (11.29)$$

In (11.29)

$$Q(\boldsymbol{y}|\beta_0, \beta_1, \eta) = \sum_{1 \leq i \leq n} y(\boldsymbol{s}_i) \left[\log\left(\frac{\kappa_i}{1-\kappa_i}\right) - \eta \sum_{j \neq i} \kappa_i\right] + \eta \sum_{1 \leq i < j \leq n} \sum y(\boldsymbol{s}_i) y(\boldsymbol{s}_j).$$

The quantity we wish to estimate is

$$E\left[\frac{1}{n}\sum_{i=1}^{n} Y(\boldsymbol{s}_i)\right] = \frac{1}{n}\sum_{i=1}^{n} E[Y(\boldsymbol{s}_i)], \tag{11.30}$$

which we believe should be similar, but not equal, to $(1/n)\sum_{i=1}^{n}\kappa_i$. To estimate the component $E[Y(\boldsymbol{s}_i)]$ we could consider approximating its definition,

$$E[Y(\boldsymbol{s}_i)] = \sum_{y(\boldsymbol{s}_i) in \{0,1\}} y(\boldsymbol{s}_i)\, p(y(\boldsymbol{s}_i)) = P[Y(\boldsymbol{s}_i) = 1]. \tag{11.31}$$

Our problem is that the marginal probability mass function of $Y(\boldsymbol{s}_i)$ is,

$$p(y(\boldsymbol{s}_i)) = \sum_{\Omega - i} p(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)),$$

where Ω_{-i} is the $(n-1)-$fold Cartesian product of $\{0,1\}$. If $k = 30$, the $k \times k$ spatial lattice of our problem has $n = 900$ locations and $|\Omega_{-i}| = 2^8 99$, a very large number. And, as already discussed, we don't know the joint distribution completely. We can accomplish our goal of estimating (11.30) using Monte Carlo. If we had values $\{\boldsymbol{y}_m^* : m = 1, \ldots, M\}$ where $\boldsymbol{y}_m^* = (y^*(\boldsymbol{s}_1), \ldots, y^*(\boldsymbol{s}_n))^T$ are simulated from the joint distribution then we could approximate, for $i = 1, \ldots, n$, $P[Y(\boldsymbol{s}_i) = 1]$ and the average of these approximations would be our estimate. We would do so using estimated values of β_0, β_1 and η. Because we have the full conditional distributions (11.28) we can simulate from the joint using a Gibbs Sampling algorithm (see Gibbs sampling for unconventional problems in Chapter 7.9.3).