12 Bayesian Procedures

12.1 Introduction

A powerful set of procedures for estimating discrete choice models has been developed within the Bayesian tradition. The breakthough concepts were introduced by Albert and Chib (1993) and McCulloch and Rossi (1994) in the context of probit, and by Allenby and Lenk (1994) and Allenby (1997) for mixed logits with normally distributed coefficients. These authors showed how the parameters of the model can be estimated without needing to calculate the choice probabilities. Their procedures provide an alternative to the classical estimation methods described in Chapter 10. Rossi *et al.* (1996), Allenby (1997), and Allenby and Rossi (1999) showed how the procedures can also be used to obtain information on individual-level parameters within a model with random taste variation. By this means, they provide a Bayesian analog to the classical procedures that we describe in Chapter 11. Variations of these procedures to accommodate other aspects of behavior have been numerous. For example, Arora *et al.* (1998) generalized the mixed logit procedure to take account of the quantity of purchases as well as brand choice in each purchase occasion. Bradlow and Fader (2001) showed how similar methods can be used to examine rankings data at an aggregate level rather than choice data at the individual level. Chib and Greenberg (1998) and Wang *et al.* (2001) developed methods for interrelated discrete responses. Chiang *et al.* (1999) examined situations where the choice set that the decision maker considers is unknown to the researcher. Train (2001) extended the Bayesian procedure for mixed logit to nonnormal distributions of coefficients, including lognormal, uniform, and triangular distributions.

The Bayesian procedures avoid two of the most prominent difficulties associated with classical procedures. First, the Bayesian procedures do not require maximization of any function. With probit and some mixed logit models (especially those with lognormal distributions), maximization of the simulated likelihood function can be difficult numerically.

Often the algorithm fails to converge for various reasons. The choice of starting values is often critical, with the algorithm converging from starting values that are close to the maximum but not from other starting values. The issue of local versus global maxima complicates the maximization further, since convergence does not guarantee that the global maximum has been attained. Second, desirable estimation properties, such as consistency and efficiency, can be attained under more relaxed conditions with Bayesian procedures than classical ones. As shown in Chapter 10, maximum simulated likelihood is consistent only if the number of draws used in simulation is considered to rise with sample size; and efficiency is attained only if the number of draws rises faster than the square root of sample size. In contrast, the Bayesian estimators that we describe are consistent for a fixed number of draws used in simulation and are efficient if the number of draws rises at any rate with sample size.

These advantages come at a price, of course. For researchers who are trained in a classical perspective, the learning curve can be steep. Numerous interrelated techniques and concepts must be assimilated before the power of them becomes clear. I can assure the reader, however, that the effort is worthwhile. Another cost of the Bayesian procedures is more fundamental. To simulate relevant statistics that are defined over a distribution, the Bayesian procedures use an iterative process that converges, with a sufficient number of iterations, to draws from that distribution. This convergence is different from the convergence to a maximum that is needed for classical procedures and involves its own set of difficulties. The researcher cannot easily determine whether convergence has actually been achieved. Thus, the Bayesian procedures trade the difficulties of convergence to a maximum for the difficulties associated with this different kind of convergence. The researcher will need to decide, in a particular setting, which type of convergence is less burdensome.

For some behavioral models and distributional specifications, Bayesian procedures are far faster and, after the initial learning that a classicist needs, are more straightforward from a programming perspective than classical procedures. For other models, the classical procedures are easier. We will explore the relative speed of Bayesian and classical procedures in the sections to follow. The differences can be readily categorized, through an understanding of how the two sets of procedures operate. The researcher can use this understanding in deciding which procedure to use in a particular setting.

Two important notes are required before proceeding. First, the Bayesian procedures, and the term "hierarchical Bayes" that is often used in the context of discrete choice models, refer to an estimation method, not a behavioral model. Probit, mixed logit, or any other model

that the researcher specifies can, in principle, be estimated by either classical or Bayesian procedures. Second, the Bayesian perspective from which these procedures arise provides a rich and intellectually satisfying paradigm for inference and decision making. Nevertheless, a researcher who is uninterested in the Bayesian perspective can still benefit from Bayesian procedures: the use of Bayesian procedures does not necessitate that the researcher adopt a Bayesian perspective on statistics. As we will show, the Bayesian procedures provide an estimator whose properties can be examined and interpreted in purely classical ways. Under certain conditions, the estimator that results from the Bayesian procedures is asymptotically equivalent to the maximum likelihood estimator. The researcher can therefore use Bayesian procedures to obtain parameter estimates and then interpret them the same as if they were maximum likelihood estimates. A highlight of the Bayesian procedures is that the results can be interpreted from both perspectives simultaneously, drawing on the insights afforded by each tradition. This dual interpretation parallels that of the classical procedures, whose results can be transformed for Bayesian interpretation as described by Geweke (1989). In short, the researcher's statistical perspective need not dictate her choice of procedure.

In the sections that follow, we provide an overview of Bayesian concepts in general, introducing the prior and posterior distributions. We then show how the mean of the posterior distribution can be interpreted from a classical perspective as being asymptotically equivalent to the maximum of the likelihood function. Next we address the numerical issue of how to calculate the mean of the posterior distribution. Gibbs sampling and, more generally, the Metropolis–Hastings algorithm can be used to obtain draws from practically any posterior distribution, no matter how complex. The mean of these draws simulates the mean of the posterior and thereby constitutes the parameter estimates. The standard deviation of the draws provides the classical standard errors of the estimates. We apply the method to a mixed logit model and compare the numerical difficulty and speed of the Bayesian and classical procedures under various specifications.

12.2 Overview of Bayesian Concepts

Consider a model with parameters θ. The researcher has some initial ideas about the value of these parameters and collects data to improve this understanding. Under Bayesian analysis, the researcher's ideas about the parameters are represented by a probability distribution over all possible values that the parameters can take, where the probability represents how likely the researcher thinks it is for the parameters to take a particular value. Prior to collecting data, the researcher's ideas are based on logic,

intuition, or past analyses. These ideas are represented by a density on θ, called the prior distribution and denoted $k(\theta)$. The researcher collects data in order to improve her ideas about the value of θ. Suppose the researcher observes a sample of N independent decision makers. Let y_n denote the observed choice (or choices) of decision maker n, and let the set of observed choices for the entire sample be labeled collectively as $Y = \{y_1, \ldots, y_N\}$. Based on this sample information, the researcher changes, or updates, her ideas about θ. The updated ideas are represented by a new density on θ, labeled $K(\theta \mid Y)$ and called the posterior distribution. This posterior distribution depends on Y, since it incorporates the information that is contained in the observed sample.

The question arises: how exactly do the researcher's ideas about θ change from observing Y? That is, how does the posterior distribution $K(\theta \mid Y)$ differ from the prior distribution $k(\theta)$? There is a precise relationship between the prior and posterior distribution, established by Bayes' rule. Let $P(y_n \mid \theta)$ be the probability of outcome y_n for decision maker n. This probability is the behavioral model that relates the explanatory variables and parameters to the outcome, though the notation for the explanatory variables is omitted for simplicity. The probability of observing the sample outcomes Y is

$$L(Y \mid \theta) = \prod_{n=1}^{N} P(y_n \mid \theta).$$

This is the likelihood function (not logged) of the observed choices. Note that it is a function of the parameters θ.

Bayes' rule provides the mechanism by which the researcher improves her ideas about θ. By the rules of conditioning,

$$K(\theta \mid Y)L(Y) = L(Y \mid \theta)\, k(\theta), \tag{12.1}$$

where $L(Y)$ is the marginal probability of Y, marginal over θ:

$$L(Y) = \int L(Y \mid \theta)k(\theta)\, d\theta.$$

Both sides of equation (12.1) represent the joint probability of Y and θ, with the conditioning in opposite directions. The left-hand side is the probability of Y times the probability of θ given Y, while the right-hand side is the probability of θ times the probability of Y given θ. Rearranging, we have

$$K(\theta \mid Y) = \frac{L(Y \mid \theta)k(\theta)}{L(Y)}. \tag{12.2}$$

This equation is Bayes' rule applied to prior and posterior distributions. In general, Bayes rule links conditional and unconditional probabilities in any setting and does not imply a Bayesian perspective on statistics. Bayesian statistics arises when the unconditional probability is the prior distribution (which reflects the researcher's ideas about θ *not* conditioned on the sample information) and the conditional probability is the posterior distribution (which gives the researcher's ideas about θ conditioned on the sample information).

We can express equation (12.2) in a more compact and convenient form. The marginal probability of Y, $L(Y)$, is constant with respect to θ and, more specifically, is the integral of the numerator of (12.2). As such, $L(Y)$ is simply the normalizing constant that assures that the posterior distribution integrates to 1, as required for any proper density. Using this fact, equation (12.2) can be stated more succinctly by saying simply that the posterior distribution is proportional to the prior distribution times the likelihood function:

$$K(\theta \mid Y) \propto L(Y \mid \theta)k(\theta).$$

Intuitively, the probability that the researcher ascribes to a given value for the parameters *after* seeing the sample is the probability that she ascribes *before* seeing the sample times the probability (i.e., *likelihood*) that those parameter values would result in the observed choices.

The mean of the posterior distribution is

$$\bar{\theta} = \int \theta K(\theta \mid Y)\, d\theta. \tag{12.3}$$

This mean has importance from both a Bayesian and a classical perspective. From a Bayesian perspective, $\bar{\theta}$ is the value of θ that minimizes the expected cost of the researcher being wrong about θ, if the cost of error is quadratic in the size of the error. From a classical perspective, $\bar{\theta}$ is an estimator that has the same asymptotic sampling distribution as the maximum likelihood estimator. We explain both of these concepts in the following sections.

12.2.1. Bayesian Properties of $\bar{\theta}$

The researcher's views about θ are represented by the posterior $K(\theta \mid Y)$ after observing the sample. Suppose that the researcher were required to guess the true value of θ and would be levied a penalty for the extent to which her guess differed from the true value. More realistically, suppose that some action must be taken that depends on the value of θ, such as a manufacturer setting the price of a good when the revenues at

any price depend on the price elasticity of demand. There is a cost to taking the wrong action, such as setting price based on the belief that the price elasticity is -0.2 when the true elasticity is actually -0.3. The question becomes: what value of θ should the researcher use in these decisions in order to minimize her expected cost of being wrong, given her beliefs about θ as represented in the posterior distribution?

If the cost of being wrong is quadratic in the distance between the θ that is used in the decision and the true θ, then the optimal value of θ to use in the decision is $\bar{\theta}$. This fact can be demonstrated as follows. If the researcher uses θ_0 in her decisions when the true value is θ^*, the cost of being wrong is

$$C(\theta_0, \theta^*) = (\theta_0 - \theta^*)' B(\theta_0 - \theta^*),$$

where B is a matrix of constants. The researcher doesn't know the true value of θ, but has beliefs about its value as represented in $K(\theta \mid Y)$. The researcher can therefore calculate the expected cost of being wrong when using the value θ_0. This expected cost is

$$\begin{aligned} \text{EC}(\theta_0) &= \int C(\theta_0, \theta) K(\theta \mid Y)\, d\theta \\ &= \int (\theta_0 - \theta)' B(\theta_0 - \theta) K(\theta \mid Y)\, d\theta. \end{aligned}$$

The value of θ_0 that minimizes this expected cost is determined by differentiating $\text{EC}(\theta^0)$, setting the derivative equal to zero, and solving for θ^0. The derivative is

$$\begin{aligned} \frac{\partial \text{EC}(\theta_0)}{\partial \theta_0} &= \int \frac{(\theta_0 - \theta)' B(\theta_0 - \theta)}{\partial \theta_0} K(\theta \mid Y)\, d\theta \\ &= \int 2(\theta_0 - \theta)' B K(\theta \mid Y)\, d\theta \\ &= 2\theta_0' B \int K(\theta \mid Y)\, d\theta - 2\left(\int \theta K(\theta \mid Y)\, d\theta \right)' B \\ &= 2\theta_0' B - 2\bar{\theta}' B. \end{aligned}$$

Setting this expression to equal zero and solving for θ_0, we have

$$\begin{aligned} 2\theta_0' B - 2\bar{\theta}' B &= 0, \\ \theta_0' B &= \bar{\theta}' B, \\ \theta_0 &= \bar{\theta}. \end{aligned}$$

The mean of the posterior, $\bar{\theta}$, is the value of θ that a Bayesian researcher would optimally act upon if the cost of being wrong about θ rises quadratically with the distance to the true θ.

Zellner (1971) describes the optimal Bayesian estimator under other loss functions. While the loss function is usually assumed to be symmetric and unbounded like the quadratic, it need not be either; see, for example, Wen and Levy (2001). Importantly, Bickel and Doksum (2000) show that the correspondence that we describe in the next section between the mean of the posterior and the maximum likelihood estimator also applies to Bayesian estimators that are optimal under many other loss functions.

12.2.2. Classical Properties of $\bar{\theta}$: The Bernstein–von Mises Theorem

Classical statistics is not concerned with the researcher's beliefs and contains no notion of prior and posterior distributions. The concern of classical statistics is to determine the sampling distribution of an estimator. This distribution reflects the fact that a different sample would produce a different point estimate. The sampling distribution is the distribution of point estimates that would be obtained if many different samples were taken. Usually, the sampling distribution for an estimator cannot be derived for small samples. However, the asymptotic sampling distribution can usually be derived, which approximates the actual sampling distribution when the sample size is large enough. In classical statistics, the asymptotic sampling distribution determines the properties of the estimator, such as whether the estimator is consistent, asymptotically normal, and efficient. The variance of the asymptotic distribution provides the standard errors of the estimates and allows for hypothesis testing, the accuracy of which rises with sample size.

From a classical perspective, $\bar{\theta}$ is simply a statistic like any other statistic. Its formula, given in (12.3), exists and can be applied even if the researcher does not interpret the formula as representing the mean of a posterior distribution. The researcher can consider $K(\theta \mid Y)$ to be a function defined by equation (12.2) for any arbitrarily defined $k(\theta)$ that meets the requirements of a density. The relevant question for the classical researcher is the same as with any statistic: what is the sampling distribution of $\bar{\theta}$?

The answer to this question is given by the Bernstein–von Mises theorem. This theorem has a long provenance and takes many forms. In the nineteenth century, Laplace (1820) observed that posterior distributions start to look more and more like normal distributions as the sample size increases. Over the years, numerous versions of the observation have been demonstrated under various conditions, and its implications have been more fully explicated. See Rao (1987), Cam and Yang (1990), Lehmann and Casella (1998), and Bickel and Doksum (2000) for modern

treatments with historical notes. The theorem is named after Bernstein (1917) and von Mises (1931) because they seem to have been the first to provide a formal proof of Laplace's observation, though under restrictive assumptions that others later relaxed.

I describe the theorem as three related statements. In these statements, the information matrix, which we used extensively in Chapters 8 and 10, is important. Recall that the score of an observation is the gradient of that observation's log likelihood with respect to the parameters: $s_n = \partial \ln P(y_n \mid \theta)/\partial\theta$, where $P(y_n \mid \theta)$ is the probability of decision maker n's observed choices. The information matrix, $-\mathbf{H}$, is the negative expected derivative of the score, evaluated at the true parameters:

$$-\mathbf{H} = -E\left(\frac{\partial^2 \ln P(y_n \mid \theta^*)}{\partial\theta\, \partial\theta'}\right),$$

where the expectation is over the population. (The negative is taken so that the information matrix can be positive definite, like a covariance matrix.) Recall also that the maximum likelihood estimator has an asymptotic variance equal to $(-\mathbf{H})^{-1}/N$. That is, $\sqrt{N}(\theta^* - \hat{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$, so that $\hat{\theta} \overset{a}{\sim} N(\theta^*, (-\mathbf{H})^{-1}/N)$, where $\hat{\theta}$ is the maximum likelihood estimator.

We can now give the three statements that collectively constitute the Bernstein–von Mises theorem:

1. $\sqrt{N}(\theta - \bar{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$.
   Stated intuitively, the posterior distribution of θ converges to a normal distribution with variance $(-\mathbf{H})^{-1}/N$ as the sample size rises. In using the expression $\xrightarrow{d}$ in this context, it is important to note that the distribution that is converging is the posterior distribution of $\sqrt{N}(\theta - \bar{\theta})$ rather than the sampling distribution. In classical analysis of estimators, as in Chapter 10, the notation $\xrightarrow{d}$ is used to indicate that the sampling distribution is converging. Bayesian analysis examines the posterior rather than the sampling distribution, and the notation indicates that the posterior distribution is converging.

   The important points to recognize in this first statement are that, as sample size rises, (i) the posterior becomes normal and (ii) the variance of the posterior becomes the same as the sampling variance of the maximum likelihood estimator. These two points are relevant for the next two statements.
2. $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$.
   The mean of the posterior converges to the maximum of the likelihood function. An even stronger statement is being made.

The difference between the mean of the posterior and the maximum of the likelihood function disappears asymptotically, *even when* the difference is scaled up by $\sqrt{N}$.

This result makes intuitive sense, given the first result. Since the posterior eventually becomes normal, and the mean and maximum are the same for a normal distribution, the mean of the posterior eventually becomes the same as the maximum of the posterior. Also, the effect of the prior distribution on the posterior disappears as the sample size rises (provided of course that the prior is not zero in the neighborhood of the true value). The posterior is therefore proportional to the likelihood function for large enough sample sizes. The maximum of the likelihood function becomes the same as the maximum of the posterior, which, as stated, is also the mean. Stated succinctly: since the posterior is asymptotically normal so that its mean equals its maximum, and the posterior is proportional to the likelihood function asymptotically, the difference between $\bar{\theta}$ and $\hat{\theta}$ eventually disappears.

3. $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$.
The mean of the posterior, considered as a classical estimator, is asymptotically equivalent to the maximum likelihood estimator. That is, $\bar{\theta} \overset{a}{\sim} N(\theta^*, (-\mathbf{H})^{-1}/N)$, just like the maximum likelihood estimator. Note that since we are now talking in classical terms, the notation refers to the sampling distribution of $\bar{\theta}$, the same as it would for any estimator.

This third statement is an implication of the first two. The statistic $\sqrt{N}(\bar{\theta} - \theta^*)$ can be rewritten as

$$\sqrt{N}(\bar{\theta} - \theta^*) = \sqrt{N}(\hat{\theta} - \theta^*) + \sqrt{N}(\bar{\theta} - \hat{\theta}).$$

From statement 2, we know that $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$, so that the second term disappears asymptotically. Only the first term affects the asymptotic distribution. This first term is the defining statistic for the maximum likelihood estimator $\hat{\theta}$. We showed in Chapter 10 that $\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$. The statistic $\sqrt{N}(\bar{\theta} - \theta^*)$ therefore follows the same distribution asymptotically. Essentially, since $\bar{\theta}$ and $\hat{\theta}$ converge, their asymptotic sampling distributions are the same.

The Bernstein–von Mises theorem establishes that $\bar{\theta}$ is on the same footing, in classical terms, as $\hat{\theta}$. Instead of maximizing the likelihood function, the researcher can calculate the mean of the posterior

distribution and know that the resulting estimator is as good in classical terms as maximum likelihood.

The theorem also provides a procedure for obtaining the standard errors of the estimates. Statement 1 says that asymptotically the variance of the posterior distribution is $(-\mathbf{H})^{-1}/N$, which, by statement 3, is the asymptotic sampling variance of the estimator $\bar{\theta}$. The variance of the posterior is the asymptotic variance of the estimates. The researcher can perform estimation entirely by using moments of the posterior: the mean of the posterior provides the point estimates, and the standard deviation of the posterior provides the standard errors.

In applications, the posterior mean and the maximum of the likelihood function can differ when sample size is insufficient for the asymptotic convergence. Huber and Train (2001) found the two to be remarkably similar in their application, while Ainslie *et al.* (2001) found them to be sufficiently different to warrant consideration. When the two estimates are not similar, other grounds must be used to choose between them (if indeed a choice is necessary), since their asymptotic properties are the same.

12.3 Simulation of the Posterior Mean

To calculate the mean of the posterior distribution, simulation procedures are generally required. As stated previously, the mean is

$$\bar{\theta} = \int \theta K(\theta \mid Y)\, d\theta.$$

A simulated approximation of this integral is obtained by taking draws of θ from the posterior distribution and averaging the results. The simulated mean is

$$\check{\theta} = \frac{1}{R} \sum_{r=1}^{R} \theta^r,$$

where θ^r is the rth draw from $K(\theta \mid Y)$. The standard deviation of the posterior, which serves as the standard error of the estimates, is simulated by taking the standard deviation of the R draws.

As stated, $\bar{\theta}$ has the same asymptotic properties as the maximum likelihood estimator $\hat{\theta}$. How does the use of simulation to approximate $\bar{\theta}$ affect its properties as an estimator? For maximum simulated likelihood (MSL), we found that the number of draws used in simulation must rise faster than the square root of the sample size in order for the estimator to be asymptotically equivalent to maximum likelihood. With a fixed

number of draws, the MSL estimator is inconsistent. If the number of draws rises with sample size but at a slower rate than the square root of the sample size, then MSL is consistent but not asymptotically normal or efficient. As we will see, desirable properties of the simulated mean of the posterior (SMP) are attained with more relaxed conditions on the number of draws. In particular, the SMP estimator is consistent and asymptotically normal for a fixed number of draws and becomes efficient and equivalent to maximum likelihood if the number of draws rises at any rate with sample size.

To demonstrate these properties, we examine the normalized statistic $\sqrt{N}(\breve{\theta} - \theta^*)$. This statistic can be rewritten as

$$\sqrt{N}(\breve{\theta} - \theta^*) = \sqrt{N}(\bar{\theta} - \theta^*) + \sqrt{N}(\breve{\theta} - \bar{\theta}).$$

From statement 3 of the Bernstein–von Mises theorem, we know the limiting distribution of the first term: $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$. The central limit theorem gives us the limiting distribution of the second term. $\breve{\theta}$ is the average of R draws from a distribution with mean $\bar{\theta}$ and variance $(-\mathbf{H}^{-1})/N$. Assuming the draws are independent, the central limit theorem states that the average of these R draws is distributed with mean $\bar{\theta}$ and variance $(-\mathbf{H})^{-1}/RN$. Plugging this information into the second term, we have $\sqrt{N}(\breve{\theta} - \bar{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1}/R)$. The two terms are independent by construction, and so

$$\sqrt{N}(\breve{\theta} - \theta^*) \xrightarrow{d} N\left(0, \left(1 + \frac{1}{R}\right)(-\mathbf{H})^{-1}\right).$$

The simulated mean of the posterior is consistent and asymptotically normal for fixed R. The covariance is inflated by a factor of $1/R$ due to the simulation; however, the covariance matrix can be calculated, and so standard errors and hypothesis testing can be conducted that take into account the simulation noise.

If R rises at any rate with N, then the second term disappears asymptotically. We have

$$\sqrt{N}(\breve{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1}),$$

which is the same as for the actual (unsimulated) mean $\bar{\theta}$ and the maximum likelihood estimator $\hat{\theta}$. When R rises with N, $\breve{\theta}$ is asymptotically efficient and equivalent to maximum likelihood.

Two notes are required regarding this derivation. First, we have assumed that the draws from the posterior distribution are independent. In the sections to follow, we describe methods for drawing from the posterior that result in draws that exhibit a type of serial correlation. When

draws of this type are used, the variance of the simulated mean is inflated by more than a factor of $1/R$. The estimator is still consistent and asymptotically normal with a fixed number of nonindependent draws; its covariance is simply greater. And, if R rises with N, the extra covariance due to simulation disappears asymptotically even with nonindependent draws, such that the simulated mean is asymptotically equivalent to maximum likelihood.

Second, we have assumed that draws from the posterior distribution can be taken without needing to simulate the choice probabilities. For some models, taking a draw from the posterior requires simulating the choice probabilities on which the posterior is based. In this case, the simulated mean of the posterior involves simulation within simulation, and the formula for its asymptotic distribution is more complex. As we will see, however, for most models, including all the models that we consider in this book, draws from the posterior can be taken without simulating the choice probabilities. One of the advantages of the Bayesian procedures is that they usually avoid the need to simulate choice probabilities.

12.4 Drawing from the Posterior

Usually, the posterior distribution does not have a convenient form from which to take draws. For example, we know how to take draws easily from a joint untruncated normal distribution; however, it is rare that the posterior takes this form for the entire parameter vector. Importance sampling, which we describe in Section 9.2.7 in relation to any density, can be useful for simulating statistics over the posterior. Geweke (1992, 1997) describes the approach with respect to posteriors and provides practical guidance on appropriate selection of a proposal density. Two other methods that we described in Chapter 9 are particularly useful for taking draws from a posterior distribution: Gibbs sampling and the Metropolis–Hasting algorithm. These methods are often called Monte Carlo Markov chain, or MCMC, methods. Formally, Gibbs sampling is a special type of Metropolis–Hasting algorithm (Gelman, 1992). However, the case is so special, and so conceptually straightforward, that the term Metropolis–Hasting (MH) is usually reserved for versions that are more complex than Gibbs sampling. That is, when the MH algorithm is Gibbs sampling, it is referred to as Gibbs sampling, and when it is more complex than Gibbs sampling, it is referred to as the MH algorithm. I maintain this convention hereafter.

It will be useful for the reader to review Sections 9.2.8 and 9.2.9, which describe Gibbs sampling and the MH algorithm, since we will be using these procedures extensively in the remainder of this chapter. As

stated, the mean of the posterior is simulated by taking draws from the posterior and averaging the draws. Instead of taking draws from the multidimensional posterior for all the parameters, Gibbs sampling allows the researcher to take draws of one parameter at a time (or a subset of parameters), conditional on values of the other parameters (Casella and George, 1992). Drawing from the posterior for one parameter conditional on the others is usually much easier than drawing from the posterior for all parameters simultaneously.

In some cases, the MH algorithm is needed in conjunction with Gibbs sampling. Suppose, for example, that the posterior for one parameter conditional on the other parameters does not take a simple form. In this case, the MH algorithm can be utilized, since it is applicable to (practically) any distribution (Chib and Greenberg, 1995).

The MH algorithm is particularly useful in the context of posterior distributions because the normalizing constant for the posterior need not be calculated. Recall that the posterior is the prior times the likelihood function, divided by a normalizing constant that assures that the posterior integrates to one:

$$K(\theta \mid Y) = \frac{L(Y \mid \theta)k(\theta)}{L(Y)},$$

where $L(Y)$ is the normalizing constant

$$L(Y) = \int L(Y \mid \theta)k(\theta)\, d\theta.$$

This constant can be difficult to calculate, since it involves integration. As described in Section 9.2.9, the MH algorithm can be applied without knowing or calculating the normalizing constant of the posterior.

In summary, Gibbs sampling, combined if necessary with the MH algorithm, allows draws to be taken from the posterior of a parameter vector for essentially any model. These procedures are applied to a mixed logit model in Section 12.6. First, however, we will derive the posterior distribution for some very simple models. As we will see, these results often apply in complex models for a subset of the parameters. This fact facilitates the Gibbs sampling of these parameters.

12.5 Posteriors for the Mean and Variance of a Normal Distribution

The posterior distribution takes a very convenient form for some simple inference processes. We describe two of these situations, which, as we will see, often arise within more complex models for a subset of the

parameters. Both results relate to the normal distribution. We first consider the situation where the variance of a normal distribution is known, but the mean is not. We then turn the tables and consider the mean to be known but not the variance. Finally, combining these two situations with Gibbs sampling, we consider the situation where both the mean and variance are unknown.

12.5.1. Result A: Unknown Mean, Known Variance

We discuss the one-dimensional case first, and then generalize to multiple dimensions. Consider a random variable β that is distributed normal with unknown mean b and known variance σ. The researcher observes a sample of N realizations of the random variable, labeled $\beta_n,\ n = 1, \ldots, N$. The sample mean is $\bar{\beta} = (1/N)\sum_n \beta_n$. Suppose the researcher's prior on b is $N(\beta_0, s_0)$; that is, the researcher's prior beliefs are represented by a normal distribution with mean b_0 and variance s_0. Note that we now have two normal distributions: the distribution of β, which has mean b, and the prior distribution on this unknown mean, which has mean β_0. The prior indicates that the researcher thinks it is most likely that $b = \beta_0$ and also thinks there is a 95 percent chance that b is somewhere between $\beta_0 - 1.96\sqrt{s_0}$ and $\beta_0 + 1.96\sqrt{s_0}$. Under this prior, the posterior on b is $N(b_1, s_1)$ where

$$b_1 = \frac{\frac{1}{s_0}b_0 + \frac{N}{\sigma}\bar{\beta}}{\frac{1}{s_0} + \frac{N}{\sigma}}$$

and

$$s_1 = \frac{1}{\frac{1}{s_0} + \frac{N}{\sigma}}.$$

The posterior mean b_1 is the weighted average of the sample mean and the prior mean.

Proof: The prior is

$$k(b) = \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0}.$$

The probability of drawing β_n from $N(b, \sigma)$ is

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma},$$

and so the likelihood of the N draws is

$$\begin{aligned} L(\beta_n\ \forall n \mid b) &= \prod_n \frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{(-N\bar{s}-N(b-\bar{\beta})^2)/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} \cdot e^{-N(b-\bar{\beta})^2/2\sigma}, \end{aligned}$$

where $\bar{s} = (1/N)\sum(\beta_n - \bar{\beta})^2$ is the sample variance of the β_n's. The posterior is therefore

$$\begin{aligned} K(b \mid \beta_n\ \forall n) &\propto L(\beta_n\ \forall n \mid b)k(b) \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} \cdot e^{-N(b-\bar{\beta})^2/2\sigma} \times \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0} \\ &= m_1 e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0]}, \end{aligned}$$

where m_1 is a constant that contains all the multiplicative terms that do not depend on b. With some algebraic manipulation, we have

$$\begin{aligned} K(b \mid \beta_n\ \forall n) &\propto e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0)]} \\ &\propto e^{(b^2-2b_1b)/2s_1} \\ &\propto e^{(b-b_1)^2/2s_1}. \end{aligned}$$

The second $\propto$ removes $\bar{\beta}^2$ and b_0^2 from the exponential, since they do not depend on b and thereby only affect the normalizing constant. (Recall that $\exp(a+b) = \exp(a)\exp(b)$, so that adding and removing terms from the exponential has a multiplicative effect on $K(b \mid \beta_n\ \forall n)$.) The third $\propto$ adds $b_1\bar{\beta}$ to the exponential, which also does not depend on b. The posterior is therefore

$$K(b \mid \beta_n\ \forall n) = me^{(b-b_1)^2/2s_1},$$

where m is the normalizing constant. This formula is the normal density with mean b_1 and variance s_1.

As stated, the mean of the posterior is a weighted average of the sample mean and the prior mean. The weight on the sample mean rises as sample size rises, so that for large enough N, the prior mean becomes irrelevant.

Often a researcher will want to specify a prior that represents very little knowledge about the parameters before taking the sample. In general,

the researcher's uncertainty is reflected in the variance of the prior. A large variance means that the researcher has little idea about the value of the parameter. Stated equivalently, a prior that is nearly flat means that the researcher considers all possible values of the parameters to be equally likely. A prior that represents little information is called *diffuse*.

We can examine the effect of a diffuse prior on the posterior of b. By raising the variance of the prior, s_0, the normal prior becomes more spread out and flat. As $s_0 \to \infty$, representing an increasingly diffuse prior, the posterior approaches $N(\bar{\beta}, \sigma/N)$.

The multivariate versions of this result are similar. Consider a K-dimensional random vector $\beta \sim N(b, W)$ with known W and unknown b. The researcher observes a sample $\beta_n,\ n = 1, \ldots, N$, whose sample mean is $\bar{\beta}$. If the researcher's prior on b is diffuse (normal with an unboundedly large variance), then the posterior is $N(\bar{\beta}, W/N)$.

Taking draws from this posterior is easy. Let L be the Choleski factor of W/N. Draw K iid standard normal deviates, $\eta_i,\ i = 1, \ldots, K$, and stack them into a vector $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calculate $\tilde{b} = \bar{\beta} + L\eta$. The resulting vector $\tilde{b}$ is a draw from $N(\bar{\beta}, W/N)$.

12.5.2. Result B: Unknown Variance, Known Mean

Consider a (one-dimensional) random variable that is distributed normal with known mean b and unknown variance σ. The researcher observes a sample of N realizations, labeled $\beta_n,\ n = 1, \ldots, N$. The sample variance *around the known mean* is $\bar{s} = (1/N)\sum_n (\beta_n - b)^2$. Suppose the researcher's prior on σ is inverted gamma with degrees of freedom v_0 and scale s_0. This prior is denoted IG(v_0, s_0). The density is zero for any negative value for σ, reflecting the fact that a variance must be positive. The mode of the inverted gamma prior is $s_0 v_0/(1 + v_0)$. Under the inverted gamma prior, the posterior on σ is also inverted gamma IG(v_1, s_1), where

$$v_1 = v_0 + N,$$
$$s_1 = \frac{v_0 s_0 + N\bar{s}}{v_0 + N}.$$

Proof: An inverted gamma with v_0 degrees of freedom and scale s_0 has density

$$k(\sigma) = \frac{1}{m_0 \sigma^{(v_0+1)/2}} e^{-v_0 s_0/2\sigma},$$

where m_0 is the normalizing constant. The likelihood of the sample,

treated as a function of σ, is

$$L(\beta_n \ \forall n \mid \sigma) = \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma} = \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma}.$$

The posterior is then

$$\begin{aligned} K(\sigma \mid \beta_n \ \forall n) &\propto L(\beta_n \ \forall n \mid \sigma) k(\sigma) \\ &\propto \frac{1}{\sigma^{N/2}} e^{-N\bar{s}/2\sigma} \times \frac{1}{\sigma^{(v_0+1)/2}} e^{-v_0 s_0/2\sigma} \\ &= \frac{1}{\sigma^{(N+v_0+1)/2}} e^{-(N\bar{s}+v_0 s_0)/2\sigma} \\ &= \frac{1}{\sigma^{(v_1+1)/2}} e^{-v_1 s_1/2\sigma}, \end{aligned}$$

which is the inverted gamma density with v_1 degrees of freedom and scale s_1.

The inverted gamma prior becomes more diffuse with lower v_0. For the density to integrate to one and have a mean, v_0 must exceed 1. It is customary to set $s_0 = 1$ when specifying $v_0 \to 1$. Under this diffuse prior, the posterior becomes $\text{IG}(1 + N, (1 + N\bar{s})/(1 + N))$. The mode of this posterior is $(1 + N\bar{s})/(2 + N)$, which is approximately the sample variance $\bar{s}$ for large N.

The multivariate case is similar. The multivariate generalization of an inverted gamma distribution is the inverted Wishart distribution. The result in the multivariate case is the same as with one random variable except that the inverted gamma is replaced by the inverted Wishart.

A K-dimensional random vector $\beta \sim N(b, W)$ has known b but unknown W. A sample of size N from this distribution has variance around the known mean of $\bar{S} = (1/N)\sum_n (\beta_n - b)(\beta_n - b)'$. If the researcher's prior on W is inverted Wishart with v_0 degrees of freedom and scale matrix S_0, labeled $\text{IW}(v_0, S_0)$, then the posterior on W is $\text{IW}(v_1, S_1)$ where

$$\begin{aligned} v_1 &= v_0 + N, \\ S_1 &= \frac{v_0 S_0 + N\bar{S}}{v_0 + N}. \end{aligned}$$

The prior becomes more diffuse with lower v_0, though v_0 must exceed K in order for the prior to integrate to one and have means. With $S_0 = I$, where I is the K-dimensional identity matrix, the posterior under a diffuse prior becomes $\text{IW}(K + N, (KI + N\bar{S})/(K + N))$. Conceptually, the prior is equivalent to the researcher having a previous sample of K observations whose sample variance was I. As N rises without bound, the influence of the prior on the posterior eventually disappears.

It is easy to take draws from inverted gamma and inverted Wishart distributions. Consider first an inverted gamma IG(v_1, s_1). Draws are taken as follows:

1. Take v_1 draws from a standard normal, and label the draws $\eta_i,\ i = 1, \ldots, v_1$.
2. Divide each draw by $\sqrt{s_1}$, square the result, and take the average. That is, calculate $r = (1/v_1)\sum_i(\sqrt{1/s_1}\eta_i)^2$, which is the sample variance of v_1 draws from a normal distribution whose variance is $1/s_1$.
3. Take the inverse of r: $\tilde{s} = 1/r$ is a draw from the inverted gamma.

Draws from a K-dimensional inverted Wishart IW(v_1, S_1) are obtained as follows:

1. Take v_1 draws of K-dimensional vectors whose elements are independent standard normal deviates. Label these draws η_i, $i = 1, \ldots, v_1$.
2. Calculate the Choleski factor of the inverse of S_1, labeled L, where $\mathrm{LL}' = S_1^{-1}$.
3. Create $R = (1/v_1)\sum_i(L\eta_i)(L\eta_i)'$. Note that R is the variance of draws from a distribution with variance S_1^{-1}.
4. Take the inverse of R. The matrix $\tilde{S} = R^{-1}$ is a draw from IW(v_1, S_1).

12.5.3. Unknown Mean and Variance

Suppose that both the mean b and variance W are unknown. For neither of these parameters does the posterior take a convenient form. However, draws can easily be obtained using Gibbs sampling and results A and B. A draw of b is taken conditional on W, and then a draw of W is taken conditional on b. Result A says that the posterior for b conditional on W is normal, which is easy to draw from. Result B says that the posterior for W conditional on b is inverted Wishart, which is also easy to draw from. Iterating through numerous cycles of draws from the conditional posteriors provides, eventually, draws from the joint posterior.

12.6 Hierarchical Bayes for Mixed Logit

In this section we show how the Bayesian procedures can be used to estimate the parameters of a mixed logit model. We utilize the approach

developed by Allenby (1997), implemented by SawtoothSoftware (1999), and generalized by Train (2001). Let the utility that person n obtains from alternative j in time period t be

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

where ε_{njt} is iid extreme value and $\beta_n \sim N(b, W)$. Giving β_n a normal distribution allows us to use results A and B, which speeds estimation considerably. In the following section, we discuss the use of nonnormal distributions.

The researcher has priors on b and W. Suppose the prior on b is normal with an unboundedly large variance. Suppose that the prior on W is inverted Wishart with K degrees of freedom and scale matrix I, the K-dimensional identity matrix. Note that these are the priors used for results A and B. More flexible priors can be specified for W, using the procedures of, for example, McCulloch and Rossi (2000), though doing so makes the Gibbs sampling more complex.

A sample of N people is observed. The chosen alternatives in all time periods for person n are denoted $y_n' = \langle y_{n1}, \ldots, y_{nT} \rangle$, and the choices of the entire sample are labeled $Y = \langle y_1, \ldots, y_N \rangle$. The probability of person n's observed choices, conditional on β, is

$$L(y_n \mid \beta) = \prod_t \left(\frac{e^{\beta' x_{ny_{nt}t}}}{\sum_j e^{\beta' x_{njt}}} \right).$$

The probability *not* conditional on β is the integral of $L(y_n \mid \beta)$ over all β:

$$L(y_n \mid b, W) = \int L(y_n \mid \beta) \phi(\beta \mid b, W)\, d\beta,$$

where $\phi(\beta \mid b, W)$ is the normal density with mean b and variance W. This $L(y_n \mid b, W)$ is the mixed logit probability.

The posterior distribution of b and W is, by definition,

$$K(b, W \mid Y) \propto \prod_n L(y_n \mid b, W) k(b, W), \tag{12.4}$$

where $k(b, W)$ is the prior on b and W described earlier (i.e., normal for b times inverted Wishart for W).

It would be *possible* to draw directly from $K(b, W \mid Y)$ with the MH algorithm. However, doing so would be computationally very slow. For each iteration of the MH algorithm, it would be necessary to calculate the right-hand side of (12.4). However, the choice probability $L(y_n \mid b, W)$ is an integral without a closed form and must be approximated through simulation. Each iteration of the MH algorithm would

therefore require simulation of $L(y_n \mid b, W)$ for each n. That would be very time-consuming, and the properties of the resulting estimator would be affected by it. Recall that the properties of the simulated mean of the posterior were derived under the assumption that draws can be taken from the posterior without needing to simulate the choice probabilities. MH applied to (12.3) violates this assumption.

Drawing from $K(b, W \mid Y)$ becomes fast and simple if each β_n is considered to be a parameter along with b and W, and Gibbs sampling is used for the three sets of parameters b, W, and $\beta_n \; \forall n$. The posterior for b, W, and $\beta_n \; \forall n$ is

$$K(b, W, \beta_n \; \forall n \mid Y) \propto \prod_n L(y_n \mid \beta_n)\phi(\beta_n \mid b, W)k(b, W).$$

Draws from this posterior are obtained through Gibbs sampling. A draw of each parameter is taken, conditional on the other parameters: (1) Take a draw of b conditional on values of W and $\beta_n \; \forall n$. (2) Take a draw of W conditional on values of b and $\beta_n \; \forall n$. (3) Take a draw of $\beta_n \forall n$ conditional on values of b and W. Each of these steps is easy, as we will see. Step 1 uses result A, which gives the posterior of the mean given the variance. Step 2 uses result B, which gives the posterior of the variance given the mean. Step 3 uses an MH algorithm, but in a way that does not involve simulation within the algorithm. Each step is described in the following.

1. $b \mid W, \beta_n \; \forall n$. We condition on W and each person's β_n in this step, which means that we treat these parameters as if they were known. Result A gives us the posterior distribution of b under these conditions. The β_n's constitute a sample of N realizations from a normal distribution with unknown mean b and known variance W. Given our diffuse prior on b, the posterior on b is $N(\bar{\beta}, W/N)$, where $\bar{\beta}$ is the sample mean of the β_n's. A draw from this posterior is obtained as described in Section 12.5.1.
2. $W \mid b, \beta_n \; \forall n$. Result B gives us the posterior for W conditional on b and the β_n's. The β_n's constitute a sample from a normal distribution with known mean b and unknown variance W. Under our prior on W, the posterior on W is inverted Wishart with $K + N$ degrees of freedom and scale matrix $(KI + NS_1)/(K + N)$, where $S_1 = (1/N)\sum_n(\beta_n - b)(\beta_n - b)'$ is the sample variance of the β_n's around the known mean b. A draw from the inverted Wishart is obtained as described in Section 12.5.2.
3. $\beta_n \mid b, W$. The posterior for each person's β_n, conditional on their choices and the population mean and variance of β_n, is

$$K(\beta_b \mid b, W, y_n) \propto L(y_n \mid \beta_n)\phi(\beta_n \mid b, W). \tag{12.5}$$

There is no simple way to draw from this posterior, and so the MH algorithm is used. Note that the right-hand side of (12.5) is easy to calculate: $L(y_n \mid \beta_n)$ is a product of logits, and $\phi(\beta_n \mid b, W)$ is the normal density. The MH algorithm operates as follows:

(a) Start with a value β_n^0.
(b) Draw K independent values from a standard normal density, and stack the draws into a vector labeled η^1.
(c) Create a trial value of β_n^1 as $\tilde{\beta}_n^1 = \beta_n^0 + \rho L \eta^1$, where ρ is a scalar specified by the researcher and L is the Choleski factor of W. Note that the proposal distribution (which is labeled $g(\cdot)$ in Section 9.2.9) is specified to be normal with zero mean and variance $\rho^2 W$.
(d) Draw a standard uniform variable μ^1.
(e) Calculate the ratio

$$F = \frac{L(y_n \mid \tilde{\beta}_n^1)\phi(\tilde{\beta}_n^1 \mid b, W)}{L(y_n \mid \beta_n^0)\phi(\beta_n^0 \mid b, W)}.$$

(f) If $\mu^1 \leq F$, accept $\tilde{\beta}_n^1$ and let $\beta_n^1 = \tilde{\beta}_n^1$. If $\mu^1 > F$, reject $\tilde{\beta}_n^1$ and let $\beta_n^1 = \beta_n^0$.
(g) Repeat the process many times. For high enough t, β_n^t is a draw from the posterior.

We now know how to draw from the posterior for each parameter conditional on the other parameters. We combine the procedures into a Gibbs sampler for the three sets of parameters. Start with any initial values b^0, W^0, and $\beta_n^0 \, \forall n$. The tth iteration of the Gibbs sampler consists of these steps:

1. Draw b^t from $N(\bar{\beta}^{t-1}, W^{t-1}/N)$, where $\bar{\beta}^{t-1}$ is the mean of the β_n^{t-1}'s.
2. Draw W_t from $\text{IW}(K + N, (KI + NS^{t-1})/(K + N))$, where $S^{t-1} = \sum_n (\beta_n^{t-1} - b^t)(\beta_n^{t-1} - b^t)'/N$.
3. For each n, draw β_n^t using one iteration of the MH algorithm previously described, starting from β_n^{t-1} and using the normal density $\phi(\beta_n \mid b^t, W^t)$.

These three steps are repeated for many iterations. The resulting values converge to draws from the joint posterior of b, W, and $\beta_n \forall n$. Once the converged draws from the posterior are obtained, the mean and standard deviation of the draws can be calculated to obtain estimates and standard errors of the parameters. Note that this procedure provides information about β_n for each n, similar to the procedure described in Chapter 11 using classical estimation.

As stated, the Gibbs sampler converges, with enough iterations, to draws from the joint posterior of all the parameters. The iterations prior to convergence are often called *burn-in*. Unfortunately, it is not always easy to determine when convergence has been achieved, as emphasized by Kass *et al.* (1998). Cowles and Carlin (1996) provide a description of the various tests and diagnostics that have been proposed. For example, Gelman and Rubin (1992) suggest starting the Gibbs sampler from several different points and testing the hypothesis that the statistic of interest (in our case, the posterior mean) is the same when calculated from each of the presumably converged sequences. Sometimes convergence is fairly obvious, so that formal testing is unnecessary. During burn-in, the researcher will usually be able to see the draws trending, that is, moving toward the mass of the posterior. After convergence has been achieved, the draws tend to move around ("traverse") the posterior.

The draws from Gibbs sampling are correlated over iterations even after convergence has been achieved, since each iteration builds on the previous one. This correlation does not prevent the draws from being used for calculating the posterior mean and standard deviation, or other statistics. However, the researcher can reduce the amount of correlation among the draws by using only a portion of the draws that are obtained after convergence. For example, the researcher might retain every tenth draw and discard the others, thereby reducing the correlation among the retained draws by an order of 10. A researcher might therefore specify a total of 20,000 iterations in order to obtain 1000 draws: 10,000 for burn-in and 10,000 after convergence, of which every tenth is retained.

One issue remains. In the MH algorithm, the scalar ρ is specified by the researcher. This scalar determines the size of each jump. Usually, smaller jumps translate into more accepts, and larger jumps result in fewer accepts. However, smaller jumps mean that the MH algorithm takes more iterations to converge and embodies more serial correlation in the draws after convergence. Gelman *et al.* (1995, p. 335) have examined the optimal acceptance rate in the MH algorithm. They found that the optimal rate is about 0.44 when $K = 1$ and drops toward 0.23 as K rises. The value of ρ can be set by the researcher to achieve an acceptance rate in this neighborhood, lowering ρ to obtain a higher acceptance rate and raising it to get a lower acceptance rate.

In fact, ρ can be adjusted within the iterative process. The researcher sets the initial value of ρ. In each iteration, a trial β_n is accepted or rejected for each sampled n. If in an iteration, the acceptance rate among the N observations is above a given value (say, 0.33), then ρ is raised. If the acceptance rate is below this value, ρ is lowered. The value of ρ then moves during the iteration process to attain the specified acceptance level.

12.6.1. Succinct Restatement

Now that the Bayesian procedures have been fully described, the model and the Gibbs sampling can be stated succinctly, in the form that is used in most publications. The model is as follows.
Utility:

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$
$$\varepsilon_{njt} \text{iid extreme value},$$
$$\beta_n \sim N(b, W).$$

Observed choice:

$$y_{nt} = i \quad \text{if and only if} \quad U_{nit} > U_{njt}\ \forall j \neq i.$$

Priors:

$$k(b, W) = k(b)k(W),$$

where

$$k(b) \quad \text{is } N(b_0, S_0) \text{ with extremely large variance,}$$
$$k(W) \text{ is IW } (K, I).$$

Conditional posteriors:

$$K(\beta_n \mid b, W, y_n) \propto \prod_t \frac{e^{\beta_n' x_{ny_{nt}t}}}{\sum_j e^{\beta_n' x_{njt}}} \phi(\beta_n \mid b, W)\ \forall n,$$
$$K(b \mid W, \beta_n\ \forall n) \text{ is } N(\bar{\beta}, W/N)), \qquad \text{where } \bar{\beta} = \sum_n \beta_n / N,$$
$$K(W \mid b, \beta_n\ \forall n) \text{ is IW}\left(K + N, \frac{KI + N\bar{S}}{K + N}\right),$$
$$\text{where } \bar{S} = \textstyle\sum_n (\beta_n - b)(\beta_n - b)'/N.$$

The three conditional posteriors are called *layers* of the Gibbs sampling. The first layer for each n depends only on data for that person, rather than for the entire sample. The second and third layers do not depend on the data directly, only on the draws of β_n, which themselves depend on the data.

The Gibbs sampling for this model is fast for two reasons. First, none of the layers requires integration. In particular, the first layer utilizes a product of logit formulas for a given value of β_n. The Bayesian procedure avoids the need to calculate the mixed logit probability, utilizing instead the simple logits conditional β_n. Second, layers 2 and 3 do not utilize the data at all, since they depend only on the draws of $\beta_n\ \forall n$. Only the mean and variance of the β_n's need be calculated in these layers.

The procedure is often called *hierarchical Bayes* (HB), because there is a hierarchy of parameters. β_n are the *individual-level parameters* for person n, which describe the tastes of that person. The β_n's are distributed in the population with mean b and variance W. The parameters b and W are often called the *population-level parameters* or *hyper-parameters*. There is also a hierarchy of priors. The prior on each person's β_n is the density of β_n in the population. This prior has parameters (hyper-parameters), namely its mean b and variance W, which themselves have priors.

12.7 Case Study: Choice of Energy Supplier

We apply the Bayesian procedures to the data that were described in Chapter 11 regarding customers' choice among energy suppliers. The Bayesian estimates are compared with estimates obtained through maximum simulated likelihood (MSL).

Each of 361 customers was presented with up to 12 hypothetical choice situations. In each choice situation, four energy suppliers were described, and the respondent was asked which one he would choose if facing the choice in the real world. The suppliers were differentiated on the basis of six factors: (1) whether the supplier charged fixed prices, and if so the rate in cents per kilowatthour, (2) the length of contract in years, during which the rates were guaranteed and the customer would be required a penalty to switch to another supplier, (3) whether the supplier was the local utility, (4) whether the supplier was a well-known company other than the local utility, (5) whether the supplier charged time-of-day (TOD) rates (specified prices in each period), and (6) whether the supplier charged seasonal rates (specified prices in each season). In the experimental design, the fixed rates varied over situations, but the same prices were specified in all experiments whenever a supplier was said to charge TOD or seasonal rates. The coefficient of the dummies for TOD and seasonal rates therefore reflect the value of these rates at the specified prices. The coefficient of the fixed price indicates the value of each cent per kilowatthour.

12.7.1. Independent Normal Coefficients

A mixed logit model was estimated under the initial assumption that the coefficients are independently normally distributed in the population. That is, $\beta_n \sim N(b, W)$ with diagonal W. The population parameters are the mean and standard deviation of each coefficient. Table 12.1 gives the simulated mean of the posterior (SMP) for these parameters, along with the MSL estimates. For the Bayesian procedure, 20,000

Table 12.1. *Mixed logit model of choice among energy suppliers*

Estimates[a]		MSL	SMP	Scaled MSL
Price coeff.:	Mean	−0.976	−1.04	−1.04
		(.0370)	(.0374)	(.0396)
	St. dev.	0.230	0.253	0.246
		(.0195)	(.0169)	(.0209)
Contract coeff.:	Mean	−0.194	−0.240	−0.208
		(.0224)	(.0269)	(.0240)
	St. dev.	0.405	0.426	0.434
		(.0238)	(.0245)	(.0255)
Local coeff.:	Mean	2.24	2.41	2.40
		(.118)	(.140)	(.127)
	St. dev.	1.72	1.93	1.85
		(.122)	(.123)	(.131)
Well-known coeff.:	Mean	1.62	1.71	1.74
		(.0865)	(.100)	(.0927)
	St. dev.	1.05	1.28	1.12
		(.0849)	(.0940)	(.0910)
TOD coeff.:	Mean	−9.28	−10.0	−9.94
		(.314)	(.315)	(.337)
	St. dev.	2.00	2.51	2.14
		(.147)	(.193)	(.157)
Seasonal coeff.:	Mean	−9.50	−10.2	−10.2
		(.312)	(.310)	(.333)
	St. dev.	1.24	1.66	1.33
		(.188)	(.182)	(.201)

[a] Standard errors in parentheses.

iterations of the Gibbs sampling were performed. The first 10,000 iterations were considered burn-in, and every tenth draw was retained after convergence, for a total of 1000 draws from the posterior. The mean and standard deviation of these draws constitutes the estimates and standard errors. For MSL, the mixed logit probability was simulated with 200 Halton draws for each observation.

The two procedures provide similar results in this application. The scale of the estimates from the Bayesian procedure is somewhat larger than that for MSL. This difference indicates that the posterior is skewed, with the mean exceeding the mode. When the MSL estimates are scaled to have the same estimated mean for the price coefficient, the two sets of estimates are remarkably close, in standard errors as well as point estimates. The run time was essentially the same for each approach.

In other applications, e.g., Ainslie *et al.* (2001), the MSL and SMP estimates have differed. In general, the magnitude of differences depends on the number of observations relative to the number of parameters, as

well as the amount of variation that is contained in the observations. When the two sets of estimates differ, it means that the asymptotics are not yet operating completely (i.e., the sample size is insufficient for the asymptotic properties to be fully exhibited). The researcher might want to apply a Bayesian perspective in this case (if she is not already doing so) in order to utilize the Bayesian approach to small-sample inference. The posterior distribution contains the relevant information for Bayesian analysis with any sample size, whereas the classical perspective requires the researcher to rely on asymptotic formulas for the sampling distribution that need not be meaningful with small samples. Allenby and Rossi (1999) provide examples of the differences and the value of the Bayesian approaches and perspective.

We reestimated the model under a variety of other distributional assumptions. In the following sections, we describe how each method is implemented under these alternative assumptions. For reasons that are inherent in the methodologies, the Bayesian procedures are easier and faster for some specifications, while the classical procedures are easier and faster for others. Understanding these realms of relative convenience can assist the researcher in deciding which method to use for a particular model.

12.7.2. Multivariate Normal Coefficients

We now allow the coefficients to be correlated. That is, W is full rather than diagonal. The classical procedure is the same except that drawing from $\phi(\beta_n \mid b, W)$ for the simulation of the mixed logit probability requires creating correlation among independent draws from a random number generator. The model is parameterized in terms of the Choleski factor of W, labeled L. The draws are calculated as $\tilde{\beta}_n = b + L\eta$, where η is a draw of a K-dimensional vector of independent standard normal deviates. In terms of computation time for MSL, the main difference is that the model has far more parameters with full W than when W is diagonal: $K + K(K+1)/2$ rather than the $2K$ parameters for independent coefficients. In our case with $K = 6$, the number of parameters rises from 12 to 27. The gradient with respect to each of the new parameters takes time to calculate, and the model requires more iterations to locate the maximum over the larger-dimensioned log-likelihood function. As shown in the second line of Table 12.2, the run time nearly triples for the model with correlated coefficients, relative to independent coefficients.

With the Bayesian procedure, correlated coefficients are no harder to handle than uncorrelated ones. For full W, the inverted gamma

Table 12.2. *Run times*

	Run time (min)	
Specification	MSL	SMP
All normal, no correlations	48	53
All normal, full covariance	139	55
1 fixed, others normal, no corr.	42	112
3 lognormal, 3 normal, no corr.	69	54
All triangular, no corr.	56	206

distribution is replaced with its multivariate generalization, the inverted Wishart. Draws are obtained by the procedure in Section 12.5.2. The only extra computer time relative to independent coefficients arises in the calculation of the covariance matrix of the β_n's and its Choleski factor, rather than the standard deviations of the β_n's. This difference is trivial for typical numbers of parameters. As shown in Table 12.2, the run time for the model with full covariance among the random coefficients was essentially the same as with independent coefficients.

12.7.3. Fixed Coefficients for Some Variables

There are various reasons that the researcher might choose to specify some of the coefficients as fixed.

1. Ruud (1996) argues that a mixed logit with all random coefficients is nearly unidentified empirically, since only ratios of coefficients are economically meaningful. He recommends holding at least one coefficient fixed, particularly when the data contain only one choice situation for each decision maker.
2. In a model with alternative-specific constants, the final iid extreme value terms constitute the random portion of these constants. Allowing the coefficients of the alternative-specific dummies to be random in addition to having the final iid extreme value terms is equivalent to assuming that the constants follow a distribution that is a mixture of extreme value and whatever distribution is assumed for these coefficients. If the two distributions are similar, such as a normal and extreme value, the mixture can be unidentifiable empirically. In this case, the analyst might choose to keep the coefficients of the alternative-specific constants fixed.
3. The goal of the analysis might be to forecast substitution patterns correctly rather than to understand the distribution of

coefficients. In this case, error components can be specified that capture the correct substitution patterns while holding the coefficients of the original explanatory variables fixed (as in Brownstone and Train, 1999).

4. The willingness to pay (wtp) for an attribute is the ratio of the attribute's coefficient to the price coefficient. If the price coefficient is held fixed, the distribution of wtp is simply the scaled distribution of the attribute's coefficient. The distribution of wtp is more complex when the price coefficient varies also. Furthermore, if the usual distributions are used for the price coefficient, such as normal or lognormal, the issue arises of how to handle positive price coefficients, price coefficients that are close to zero so that the implied wtp is extremely high, and price coefficients that are extremely negative. The first of these issues is avoided with lognormals, but not the other two. The analyst might choose to hold the price coefficient fixed to avoid these problems.

In the classical approach, holding one or more coefficients fixed is very easy. The corresponding elements of W and L are simply set to zero, rather than treated as parameters. The run time is reduced, since there are fewer parameters. As indicated in the third line of Table 12.2, the run time decreased by about 12 percent with one fixed coefficient and the rest independent normal, relative to all independent normals. With correlated normals, a larger percentage reduction would occur, since the number of parameters drops more than proportionately.

In the Bayesian procedure, allowing for fixed coefficients requires the addition of a new layer of Gibbs sampling. The fixed coefficient cannot be drawn as part of the MH algorithm for the random coefficients for each person. Recall that under MH, trial draws are accepted or rejected in each iteration. If a trial draw which contains a new value of a fixed coefficient along with new values of the random coefficients is accepted for one person, but the trial draw for another person is not accepted, then the two people will have different values of the fixed coefficient, which contradicts the fact that it is fixed. Instead, the random coefficients, and the population parameters of these coefficients, must be drawn conditional on a value of the fixed coefficients; and then the fixed coefficients are drawn conditional on the values of the random coefficients. Drawing from the conditional posterior for the fixed coefficients requires an MH algorithm, in addition to the one that is used to draw the random coefficients.

To be explicit, rewrite the utility function as

$$U_{njt} = \alpha' z_{njt} + \beta_n' x_{njt} + \varepsilon_{njt}, \tag{12.6}$$

where α is a vector of fixed coefficients and β_n is random as before with mean b and variance W. The probability of the person's choice sequence given α and β_n is

$$(12.7)\quad L(y_n \mid \alpha, \beta_n) = \prod_t \frac{e^{\alpha' z_{ny_{nt}t} + \beta_n' x_{ny_{nt}t}}}{\sum_j e^{\alpha' z_{njt} + \beta_n' x_{njt}}}.$$

The conditional posteriors for Gibbs sampling are:

1. $K(\beta_n \mid \alpha, b, W) \propto L(y_n \mid \alpha, \beta_n)\phi(\beta_n \mid b, W)$. MH is used for these draws in the same way as with all normals, except that now $\alpha' z_{njt}$ is included in the logit formulas.
2. $K(b \mid W, \beta_n\ \forall n)$ is $N(\Sigma_n \beta_n/N, W/N)$. Note that α does not enter this posterior; its effect is incorporated into the draws of β_n from layer 1.
3. $K(W \mid b, \beta_n\ \forall n)$ is $\text{IW}(K + N, (KI + N\bar{S})/(K + N))$, where $\bar{S} = \Sigma_n(\beta_n - b)(\beta_n - b)'/N$. Again, α does not enter directly.
4. $K(\alpha \mid \beta_n) \propto \Pi_n L(y_n \mid \alpha, \beta_n)$, if the prior on α is essentially flat (e.g., normal with sufficiently large variance). Draws are obtained with MH on the pooled data.

Layer 4 takes as much time as layer 1, since each involves calculation of a logit formula for each observation. The Bayesian procedure with fixed and normal coefficients can therefore be expected to take about twice as much time as with all normal coefficients. As indicated in the third line of Table 12.2, this expectation is confirmed in our application.

12.7.4. Lognormals

Lognormal distributions are often specified when the analyst wants to assure that the coefficient takes the same sign for all people. There is little change in either procedure when some or all of the coefficients are distributed lognormal instead of normal. Normally distributed coefficients are drawn, and then the ones that are lognormally distributed are exponentiated when they enter utility. With all lognormals, utility is specified as

$$(12.8)\quad U_{njt} = (e^{\beta_n})' x_{njt} + \varepsilon_{njt},$$

with β_n distributed normal as before with mean b and variance W. The probability of the person's choice sequence given β_n is

$$(12.9)\quad L(y_n \mid \alpha, \beta_n) = \prod_t \frac{e^{(e^{\beta_n})' x_{ny_{nt}t}}}{\sum_j e^{(e^{\beta_n})' x_{njt}}}.$$

With this one change, the rest of the steps are the same with both procedures. In the classical approach, however, locating the maximum of the likelihood function is considerably more difficult with lognormal coefficients than with normal ones. Often the numerical maximization procedures fail to find an increase after a number of iterations. Or a "maximum" is found and yet the Hessian is singular at that point. It is often necessary to specify starting values that are close to the maximum. And the fact that the iterations can fail at most starting values makes it difficult to determine whether a maximum is local or global. The Bayesian procedure does not encounter these difficulties, since it does not search for the maximum. The Gibbs sampling seems to converge a bit more slowly, but not appreciably so. As indicated in Table 12.2, the run time for the classical approach rose nearly 50 percent with lognormal relative to normals (due to more iterations being needed), while the Bayesian procedure took about the same amount of time with each. This comparison is generous to the classical approach, since convergence at a maximum was achieved in this application, while in many other applications we have not been able to obtain convergence with lognormals or have done so only after considerable time was spent finding successful starting values.

12.7.5. Triangulars

Normal and lognormal distributions allow coefficients of unlimited magnitude. In some situations, the analyst might want to assure that the coefficients for all people remain within a reasonable range. This goal is accomplished by specifying distributions that have bounded support, such as uniform, truncated normal, and triangular distributions. In the classical approach, these distributions are easy to handle. The only change occurs in the line of code that creates the random draws from the distributions. For example, the density of a triangular distribution with mean b and spread s is zero beyond the range $(b - s, b + s)$, rises linearly from $b - s$ to b, and drops linearly to $b + s$. A draw is created as $\beta_n = b + s(\sqrt{2\mu} - 1)$ if $\mu < 0.5$ and $= b + s(1 - \sqrt{2(1 - \mu)})$ otherwise, where μ is a draw from a standard uniform. Given draws of β_n, the calculation of the simulated probability and the maximization of the likelihood function are the same as with draws from a normal. Experience indicates that estimation of the parameters of uniform, truncated normal, and triangular distributions takes about the same number of iterations as for normals. The last line of Table 12.2 reflects this experience.

With the Bayesian approach, the change to nonnormal distributions is far more complicated. With normally distributed coefficients, the conditional posteriors for the population moments are very convenient:

normal for the mean and inverted Wishart for the variance. Most other distributions do not give such convenient posteriors. Usually, an MH algorithm is needed for the population parameters, in addition to the MH algorithm for the customer-level β_n's. This addition adds considerably to computation time. The issue is exacerbated for distributions with bounded support, since, as we see in the following, the MH algorithm can be expected to converge slowly for these distributions.

With independent triangular distributions for all coefficients with mean and spread vectors b and s, and flat priors on each, the conditional posteriors are:

1. $K(\beta_n \mid b, s) \propto L(y_n \mid \beta_n)h(\beta_n \mid b, s)$, where h is the triangular density. Draws are obtained through MH, separately for each person. This step is the same as with independent normals except that the density for β_n is changed.
2. $K(b, s \mid \beta_n) \propto \prod_n h(\beta_n \mid b, s)$ when the priors on b and s are essentially flat. Draws are obtained through MH on the β_n's for all people.

Because of the bounded support of the distribution, the algorithm is exceedingly slow to converge. Consider, for example, the spread of the distribution. In the first layer, draws of β_n that are outside the range $(b - s, b + s)$ from the second layer are necessarily rejected. And in the second layer, draws of b and s that create a range $(b - s, b + s)$ that does not cover all the β_n's from the first layer are necessarily rejected. It is therefore difficult for the range to grow narrower from one iteration to the next. For example, if the range is 2 to 4 in one iteration of the first layer, then the next iteration will result in values of β_n between 2 and 4 and will usually cover most of the range if the sample size is sufficiently large. In the next draw of b and s, any draw that does not cover the range of the β_n's (which is nearly 2 to 4) will be rejected. There is indeed some room for play, since the β_n's will not cover the entire range from 2 to 4. The algorithm converges, but in our application we found that far more iterations were needed to achieve a semblance of convergence, compared with normal distributions. The run time rose by a factor of four as a result.

12.7.6. Summary of Results

For normal distributions with full covariance matrices, and for transformations of normals that can be expressed in the utility function, such as exponentiating to represent lognormal distributions, the Bayesian approach seems to be very attractive computationally. Fixed coefficients add a layer of conditioning to the Bayesian approach that

doubles its run time. In contrast, the classical approach becomes faster for each coefficient that is fixed instead of random, because there are fewer parameters to estimate. For distributions with bounded support, like triangulars, the Bayesian approach is very slow, while the classical approach handles these distributions as quickly as normals.

These comparisons relate to mixed logits only. Other behavioral models can be expected to have different relative run times for the two approaches. The comparison with mixed logit elucidates the issues that arise in implementing each method. Understanding these issues assists the researcher in specifying the model and method that are most appropriate and convenient for the choice situation.

12.8 Bayesian Procedures for Probit Models

Bayesian procedures can be applied to probit models. In fact, the methods are even faster for probit models than for mixed logits. The procedure is described by Albert and Chib (1993), McCulloch and Rossi (1994), Allenby and Rossi (1999), and McCulloch and Rossi (2000). The method differs in a critical way from the procedure for mixed logits. In particular, for a probit model, the probability of each person's choices conditional on the coefficients of the variables, which is the analog to $L(y_n \mid \beta_n)$ for logit, is not a closed form. Procedures that utilize this probability, as in the first layer of Gibbs sampling for mixed logit, cannot be readily applied to probit. Instead, Gibbs sampling for probits is accomplished by considering the utilities of the alternatives, U_{njt}, to be parameters themselves. The conditional posterior for each U_{njt} is truncated normal, which is easy to draw from. The layers for the Gibbs sampling are as follows:

1. Draw b conditional on W and $\beta_n \ \forall n$.
2. Draw W conditional on b and $\beta_n \ \forall n$. These two layers are the same as for mixed logit.
3. For each n, draw β_n conditional on $U_{njt} \ \forall j, t$. These draws are obtained by recognizing that, given the value of utility, the function $U_{njt} = \beta_n x_{njt} + \varepsilon_{njt}$ is a regression of x_{njt} on U_{njt}. Bayesian posteriors for regression coefficients and normally distributed errors have been derived (similar to our results A and B) and are easy to draw from.
4. For each n, i, t, draw U_{nit} conditional on β_n and the value of U_{njt} for each $j \neq i$. As stated earlier, the conditional posterior for each U_{nit} is a univariate truncated normal, which is easy to draw from with the procedure given in Section 9.2.4.

Details are provided in the cited articles.

Bolduc *et al.* (1997) compared the Bayesian method with MSL and found the Bayesian procedure to require about half as much computer time as MSL with random draws. If Halton draws had been used, it seems that MSL would have been faster for the same level of accuracy, since fewer than half as many draws would be needed. The Bayesian procedure for probit relies on all random terms being normally distributed. However, the concept of treating the utilities as parameters can be generalized for other distributions, giving a Bayesian procedure for mixed probits.

Bayesian procedures can be developed in some form or another for essentially any behavioral model. In many cases, they provide large computational advantages over classical procedures. Examples include the dynamic discrete choice models of Imai *et al.* (2001), the joint models of the timing and quantity of purchases due to Boatwright *et al.* (2001), and Brownstone's (2001) mixtures of distinct discrete choice models. The power of these procedures, and especially the potential for cross-fertilization with classical methods, create a bright outlook for the field.