used for results A and B above. 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

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

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 above (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 M-H algorithm. However, doing so would be computationally very slow. For each iteration of the M-H 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 for and must be approximated through simulation. Each iteration of the M-H algorithm would therefore require simulation of $L(y_n \mid b, W)$ for each n. As well as being very time consuming, the properties of the resulting estimator would be affected by this simulation within the M-H algorithm. 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. M-H 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 a M-H algorithm, but in a way that does not involve simulation within the algorithm. Each step is described below.

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

$$K(\beta_b \mid b, W, y_n) \propto L(y_n \mid \beta_n)\phi(\beta_n \mid b, W) \qquad (12.5)$$

   There is no simple way to draw from this posterior, and so the M-H 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 M-H 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 t-th 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 $IW(K+N, (KI+NS^{t-1})/(K+N))$ 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 M-H algorithm described above, 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 such 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 M-H 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 means that the M-H 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 M-H algo-

rithm. They found that the optimal rate is about 0.44 when $K = 1$ and drops toward .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, .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:

$$
\begin{aligned}
&\text{Utility:} \\
U_{njt} &= \beta_n' x_{njt} + \varepsilon_{njt} \\
\varepsilon_{njt} &\ \text{iid} \ \text{ extreme value} \\
\beta_n &\sim N(b, W) \\
&\text{Observed choice:} \\
y_{nt} &= i \text{ if and only if } U_{nit} > U_{njt}\ \forall j \neq i \\
&\text{Priors:} \\
k(b, W) &= k(b)k(W) \\
&\text{where} \\
k(b) &\ \text{is} \ N(b_0, S_0) \text{ with extremely large variance} \\
k(W) &\ \text{is} \ IW(K, I)
\end{aligned}
$$

The conditional posteriors are:

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

$$\text{where } \bar{S} = \sum_n (\beta_n - b)(\beta_n - b)'/N$$

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 β_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 is 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 kilowatt-hour (kWh), (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 at specified prices in each period, and (6) whether the supplier charged seasonal rates at 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 kWh.

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 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. 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., sample size is

Table 12.1: Mixed Logit Model of Choice Among Energy Suppliers

Estimates (se's in parens.)		MSL	SMP	Scaled MSL
Price coef:	mean	-0.976	-1.04	-1.04
		(.0370)	(.0374)	(.0396)
	st dev	.230	.253	.246
		(.0195)	(.0169)	(.0209)
Contract coef:	mean	-0.194	-0.240	-0.208
		(.0224)	(.0269)	(.0240)
	st dev	.405	.426	.434
		(.0238)	(.0245)	(.0255)
Local coef:	mean	2.24	2.41	2.40
		(.118)	(.140)	(.127)
	st dev	1.72	1.93	1.85
		(.122)	(.123)	(.131)
Well-known coef:	mean	1.62	1.71	1.74
		(.0865)	(.100)	(.0927)
	st dev	1.05	1.28	1.12
		(.0849)	(.0940)	(.0910)
TOD coef:	mean	-9.28	-10.0	-9.94
		(.314)	(.315)	(.337)
	st dev	2.00	2.51	2.14
		(.147)	(.193)	(.157)
Seasonal coef:	mean	-9.50	-10.2	-10.2
		(.312)	(.310)	(.333)
	st dev	1.24	1.66	1.33
		(.188)	(.182)	(.201)

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 re-estimated 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 to the methodologies, the Bayesian procedures are easier and faster for some types of specifications, while the classical procedures are easier and faster for other specifications. 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 dis-

Table 12.2: Run-Times in minutes

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

tribution 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, 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 such 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 coefficient fixed is very easy. The corresponding elements of W and L are simply set to zero, rather than treated as parameters. Run time is reduced since there are fewer parameters. As indicated in the third line of Table 12.2, 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 percent 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 M-H algorithm for the random coefficients for each person. Recall that under M-H, 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 a M-H algorithm, in addition to the M-H algorithm 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

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

The conditional posteriors for Gibbs sampling are:

1. $K(\beta_n \mid \alpha, b, W) \propto L(y_n \mid \alpha, \beta_n)\phi(\beta_n \mid b, W)$. M-H 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 $N(\sum_n \beta_n/N, W/N))$. 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 $IW(K + N, (KI + N\bar{S})/(K + N))$ where $\bar{S} = \sum_n (\beta_n - b)(\beta_n - b)'/N$. Again, α does not enter directly.

4. $K(\alpha \mid \beta_n) \propto \prod_n L(y_n \mid \alpha, \beta_n)$, if the prior on α is essentially flat (e.g., normal with sufficiently large variance.) Draws are obtained with M-H 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. Little changes 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

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

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

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

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 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, 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 non-normal distributions is far more complicated. With normally distributed coefficients, the conditional posterior 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, a M-H algorithm is needed for the population parameters, in addition to the M-H 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 below, the M-H 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 M-H, 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 M-H 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 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. Run time rose by a factor of four as a result.

12.7.6 Summary of results

For normal distributions with full covariance matrixes, 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 is 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 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 condition 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 utility of each alternative, 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:

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 above, 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, Fortin and Gordon (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 Ching, Imai and Jain (2001), the joint models of the timing and quantity of purchases of Boatwright, Borle and Kadane (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.

Bibliography

Adamowicz, W. (1994), 'Habit formation and variety seeking in a discrete choice model of recreation demand', *Journal of Agricultural and Resource Economics* **19**, 19–31.

Ainslie, A., R. Andrews and I. Currim (2001), 'An empirical comparison of logit choice models with discrete vs. continuous representation of heterogeneity', Working Paper, Department of Business Administration, University of Delaware.

Albert, J. and S. Chib (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* **88**, 669–679.

Allenby, G. (1997), 'An introduction to hierarchical bayesian modeling', Tutorial notes, Advanced Research Techniques Forum, American Marketing Association.

Allenby, G. and P. Lenk (1994), 'Modeling household purchase behavior with logistic normal regression', *Journal of the American Statistical Association* **89**, 1218–1231.

Allenby, G. and P. Rossi (1999), 'Marketing models of consumer heterogeneity', *Journal of Econometrics* **89**, 57–78.

Amemiya, T. (1978), 'On two-step estimation of multivariate logit models', *Journal of Econometrics* **8**, 13–21.

Arora, N., G. Allenby and J. Ginter (1998), 'A hierachical bayes model of primary and secondary demand', *Marketing Science* **17**, 29–44.

Beggs, S., S. Cardell and J. Hausman (1981), 'Assessing the potential demand for electric cars', *Journal of Econometrics* **16**, 1–19.

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Princeton NJ.

Ben-Akiva, M. (1972), The Structure of Travel Demand Models, PhD thesis, MIT.

Ben-Akiva, M. and B. Francois (1983), 'Mu-homogenous generalized extreme value model', Working Paper, Department of Civil Engineering, MIT.

Ben-Akiva, M. and D. Bolduc (1996), 'Multinomial probit with a logit kernel and a general parametric specification of the covariance structure', Working Paper, Department of Civil Engineering, MIT.

Ben-Akiva, M., D. Bolduc and J. Walker (2001), 'Specification, estimation and identification of the logit kernel (or continuous mixed logit) model', Working Paper, Department of Civil Engineering, MIT.

Ben-Akiva, M., D. Bolduc and M. Bradley (1993), 'Estimation of travel model choice models with randomly distributed values of time', *Transportation Research Record* **1413**, 88–97.

Ben-Akiva, M. and M. Bierlaire (1999), Discrete choice methods and their applications in short term travel decisions, *in* R.Hall, ed., 'The Handbook of Transportation Science', Kluwer, Dordrecht, the Netherlands, pp. 5–33.

Ben-Akiva, M. and S. Lerman (1985), *Discrete Choice Analysis : Theory and Application to Travel Demand*, MIT Press, Cambridge, Massahusetts.

Ben-Akiva, M. and T. Morikawa (1990), 'Estimation of switching models from revealed preferences and stated intentions', *Transportation Research A* **24**, 485–495.

Berkovec, J. and S. Stern (1991), 'Job exit behavior of older men', *Econometrica* **59**, 189–210.

Berndt, E., B. Hall, R. Hall and J. Hausman (1974), 'Estimation and inference in nonlinear structural models', *Annals of Economic and Social Measurement* **3/4**, 653–665.

Bernstein, S. (1917), *Calcul des probabilites*.

Berry, S. (1994), 'Estimating discrete choice models of product differentiation', *RAND Journal of Economics* **25**, 242–262.

Berry, S., J. Levinsohn and A. Pakes (1995), 'Automobile prices in market equilibrium', *Econometrica* **63**, 841–889.

Bhat, C. (1995), 'A heteroscedastic extreme value model of intercity mode choice', *Transportation Research B* **29**, 471–483.

Bhat, C. (1997), 'Covariance heterogeneity in nested logit models: Econometric structure and application to intercity travel', *Transportation Research B* **31**, 11–21.

Bhat, C. (1998*a*), 'Accommodating variations in responsiveness to level-of-service variables in travel mode choice models', *Transportation Research A* **32**, 455–507.

Bhat, C. (1998*b*), 'An analysis of travel mode and departure time choice for urban shopping trips', *Transportation Research B* **32**, 361–371.

Bhat, C. (1999), 'An analysis of evening commute stop-making behavior using repeated choice observation from a multi-day survey', *Transportation Research B* **33**, 495–510.

Bhat, C. (2000), 'Incorporating observed and unobserved heterogeneity in urban work mode choice modeling', *Transportation Science* **34**, 228–238.

Bhat, C. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model', *Transportation Research B* **35**, 677–693.

Bhat, C. (forthcoming), 'Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences', *Transportation Research* .

Bhat, C. and S. Castelar (2001), 'A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the san francisco bay area', *Transportation Research* .

Bickel, P. and K. Doksum (2000), *Mathematical Statistics: Basic ideas and Selected Topics*, Vol. 1, Prentice Hall, Upper Saddle River, NJ.

Bierlaire, M. (1998), Discrete choice models, *in* m. Labbe, G.Laporte, K.Tanczos and P.Toint, eds, 'Operations Research and Decision Aid Methodologies in Traffic and Transportation Management', Spinger Verlag, Heidelberg, Germany, pp. 203–227.

Boatwright, P., S. Borle and J. Kadane (2001), 'A model of the joint distribution of purchase quantity and timing', Conference Presentation, Bayesian Applications and Methods in Marketing Conference, Ohio State University.

Bolduc, D. (1992), 'Generalized autoregressive errors: the multinomial probit model', *Transportation Research B* **26**, 155–170.

Bolduc, D. (1993), 'Maximum simulated likelihood estimation of mnp models using the ghk probability simulation with analytic derivatives', Working Paper, Department d' Economique, Universite Laval, Quebec.

Bolduc, D. (1999), 'A practical technique to estimate multinomial probit models in transportation', *Transportation Research B* **33**, 63–79.

Bolduc, D., B. Fortin and M. Fournier (1996), 'The impact of incentive policies on the practice location of doctors: A multinomial probit analysis', *Journal of Labor Economics* **14**, 703–732.

Bolduc, D., B. Fortin and S. Gordon (1997), 'Multinomial probit estimation of spatially interdependent choices: An empirical comparison of two new techniques', *International Regional Science Review* **20**, 77–101.

Borsch-Supan, A. and V. Hajivassiliou (1993), 'Smooth unbiased multivariate probablility simulation for maximum likelihood estimation of limited dependent variable models', *Journal of Econometrics* **58**, 347–368.

Borsch-Supan, A., V. Hajivassiliou, L. Kotlikoff and J. Morris (1991), Health, children, and elderly living arrangements: A multiperiod

multinomial probit model with unobserved heterogeneity and autocorrelated errors, *in* D.Wise, ed., 'Topics in the Economics of Aging', University of Chicago Press, Chicago, IL.

Boyd, J. and J. Mellman (1980), 'The effect of fuel economy standards on the u.s. automotive market: a hedonic demand analysis', *Transportation Research A* **14**, 367–378.

Braatan and Weller (1979), 'An improved low-discrepancy sequence for multi-dimensional quasi-monte carlo integration', *Journal of Computational Physics* **33**, 249–258.

Bradley, M. and Andrew Daly (1994), 'Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data', *Transportation* **21**, 167–184.

Bradlow, E. and P. Fader (2001), 'A bayesian lifetime model for the 'hot 100' billboard songs', Working Paper, The Wharton School, University of Pennsylvania.

Brownstone, D. (2001), Discrete choice modeling for transportation, *in* D.Hensher, ed., 'Travel Behavior Research: The Leading Edge', Elsevier, Oxford, UK, pp. 97–124.

Brownstone, D., D. Bunch and K. Train (2000), 'Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles', *Transportation Research B* **34**, 315–338.

Brownstone, D. and K. Small (1989), 'Efficient estimation of nested logit model', *Journal of Business and Economic Statistics* **7**, 67–74.

Brownstone, D. and K. Train (1999), 'Forecasting new product penetration with flexible substitution patterns', *Journal of Econometrics* **89**, 109–129.

Bunch, D. (1991), 'Estimability in the multinomial probit model', *Transportation Research B* **25**, 1–12.

Bunch, D. and R. Kitamura (1989), 'Multinomial probit estimation revisited: Testing new algorithms and evaluation of alternative model specification of household car ownership', Transportation Research Group Report UCD-TRG-RR-4, University of California, Davis.

Butler, J. and R. Moffitt (1982), 'A computationally efficient quadrature procedure for the one factor multinomial probit model', *Econometrica* **50**, 761–764.

Cai, Y., I. Deilami and K. Train (1998), 'Customer retention in a competitive power market: Analysis of a 'double-bounded plus follow-ups' questionnaire', *The Energy Journal* **19**, 191–215.

Cam, L. Le and G. Yang (1990), *Asymptotics in Statistics*, Springer-Verlag, New York.

Cameron, T. (1988), 'A new paradigm for valuing non-market goods using referendum data: Maximum likelihood estimation by censored logistic regression', *Journal of Environmental Economics and Management* **15**, 355–379.

Cameron, T. and J. Quiggin (1994), 'Estimation using contingent valuation data from a 'dichotomous choice with follow-up' questionnaire', *Journal of Environmental Economics and Management* **27**, 218–234.

Cameron, T. and M. James (1987), 'Efficient estimation methods for closed-ended contingent valuation survey data', *Review of Economics and Statistics* **69**, 269–276.

Cardell, S. and F. Dunbar (1980), 'Measuring the societal impacts of automobile downsizing', *Tranportation Research A* **14**, 423–434.

Carneiro, P., J. Heckman and E. Vytlacil (2001), 'Estimating the returns to education when it varies among individuals', Working Paper, Department of Economics, University of Chicago.

Casella, G. and E. George (1992), 'Explaining the gibbs sampler', *The American Statistician* **46**, 167–174.

Chapman, R. and R. Staelin (1982), 'Exploiting rank ordered choice set data within the stochastic utility model', *Journal of Marketing Research* **14**, 288–301.

Chesher, A. and J. Santos-Silva (2002), 'Taste variation in discrete choice models', *Review of Economic Studies* **69**, 62–78.

Chiang, J., S. Chib and C. Narasimhan (1999), 'Markov chain monte carlo and models of consideration set and parameter heterogeneity', *Journal of Econometrics* **89**, 223–248.

Chib, S. and E. Greenberg (1995), 'Understanding the metropolis - hastings algorithm', *American Statistician* **49**, 327–335.

Chib, S. and E. Greenberg (1996), 'Markov chain monte carlo simulation methods in econometrics', *Econometric Theory* **12**, 409–431.

Chib, S. and E. Greenberg (1998), 'Analysis of multivariate probit models', *Biometrika* **85**, 347–361.

Ching, A., S. Imai and N. Jain (2001), 'Bayesian estimation of dynamics discrete choice models', Conference Presentation, Bayesian Applications and Methods in Marketing Conference, Ohio State University.

Chintagunta, P., D. Jain and N. Vilcassim (1991), 'Investigating heterogeneity in brand preference in logit models for panel data', *Journal of Marketing Research* **28**, 417–428.

Chipman, J. (1960), 'The foundations of utility', *Econometrica* **28**, 193–224.

Chu, C. (1981), Structural Issues and Sources of Bias in Residential Location and Travel Choice Models, PhD thesis, Northwestern University.

Chu, C. (1989), 'A paired combinational logit model for travel demand analysis', *Proceedings of Fifth World Conference on Transportation Research* **4**, 295–309.

Clark, C. (1961), 'The greatest of a finite set of random variables', *Operations Research* **9**, 145–162.

Cosslett, S. (1981), Efficient estimation of discrete choice models, *in* C.Manski and D.McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT Press, Cambridge, MA.

Cowles, M. and B. Carlin (1996), 'Markov chain monte carlo convergence diagnostics: A comparative review', *Journal of American Statistical Association* **91**, 883–904.

Daganzo, C. (1979), *Multinomial Probit: The Theory and Its Application to Demand Forecasting*, Academic Press, New York.

Daganzo, C., F. Bouthelier and Y. Sheffi (1977), 'Multinomial probit and qualitative choice: A computationally efficient algorithm', *Transportation Science* **11**, 338–358.

Dagsvik, J. (1994), 'Discrete and continuous choice max-stable processes and independence from irrelevant alternatives', *Econometrica* **62**, 1179–205.

Daly, A. (1987), 'Estimating 'tree' logit models', *Transportation Research B* **21**, 251–267.

Daly, A. and S. Zachary (1978), Improved multiple choice models, *in* D.Hensher and M.Dalvi, eds, 'Determinants of Travel Choice', Saxon House, Sussex.

Debreu, G. (1960), 'Review of r.d. luce individual choice behavior', *American Economic Review* **50**, 186–88.

DeSarbo, W., V. Ramaswamy and S. Cohen (1995), 'Market segmentation with choice-based conjoint analysis', *Marketing Letters* **6**, 137–147.

Desvousges, W., S. Waters and K. Train (1996), 'Potential economic losses associated with recreational services in the upper clark fork river basin', Report, Triangle Economic Research, Durham, NC.

Dubin, J. and D. McFadden (1984), 'An econometric analysis of residential electric appliance holdings and consumption', *Econometrica* **52**, 345–362.

Eckstein, Z. and K. Wolpin (1989), 'The specification and estimation of dynamic stochastic discrete choice models: A survey', *Journal of Human resources* **24**, 562–598.

Elrod, T. and M. Keane (1995), 'A factor analytic probit model for representing the market structure in panel data', *Journal of Marketing Research* **32**, 1–16.

Erdem, T. (1996), 'A dynamic analysis of market structure based on panel data', *Marketing Science* **15**, 359–378.

Forinash, C. and F. Koppelman (1993), 'Application and interpretation of nested logit models of intercity mode choice', *Transportation Research Record* **1413**, 98–106.

Gelman, A. (1992), 'Iterative and non-iterative simulation algorithms', *Computing Science and Statistics (Interface Proceedings)* **24**, 433–438.

Gelman, A. and D. Rubin (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Sciences* **7**, 457–511.

Gelman, A., J. Carlin, H. Stern and D. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, Suffolk.

Geman, S. and D. Geman (1984), 'Stochastic relaxation gibbs distributions and the bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geweke, J. (1988), 'Antithetic acceleration of monte carlo integration in bayesian inference', *Journal of Econometrics* **38**, 73–89.

Geweke, J. (1989), 'Bayesian inference in econometric models using monte carlo integration', *Econometrica* **57**, 1317–1339.

Geweke, J. (1991), 'Efficient simulation from the multivariate normal and student -t distributions subject to linear constraints', *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface* pp. 571–578.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *in* J.Bernardo, J.Berger, A.Dawid and F.Smith, eds, 'Bayesian Statistics', Oxford University Press, New York, pp. 169–193.

Geweke, J. (1996), Monte carlo simulation and numerical integration, *in* D.Kendrick and J.Rust, eds, 'Handbook of Computational Economics', Elsevier Science, Amsterdam, pp. 731–800.

Geweke, J. (1997), Posterior simulators in econometrics, *in* D.Kreps and K.Wallis, eds, 'Advance Economics and Econometric Theory and Applications', Cambridge University Press, New York.