A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit

Kenneth Train
Department of Economics
University of California,Berkeley

June 18, 2001

1 Introduction

Mixed logit is a flexible discrete choice model that allows for random coefficients and/or error components that induce correlation over alternatives and time. Procedures for estimating mixed logits have been developed within both the classical (e.g., Revelt and Train, 1998) and Bayesian traditions (Sawtooth Software, 1999.) Asymptotically, the two procedures provide the same information, and Huber and Train (2001) found that the two methods provide very similar results on their typical sample. The relative convenience of the two methods, in terms of programming and computation time, depends on the specification of the model. The purpose of this paper is to elucidate these realms of relative convenience.

Analogous concepts apply for probit models. The classical approach is usually implemented with the GHK probit simulator developed by Geweke (1989), Hajivassiliou (1990), and Keane (1990). A Bayesian approach has been developed by Albert and Chib (1993), McCulloch and Rossi (1994), and Allenby and Rossi (1999). A run-time comparison of the two procedures has been conducted by Bolduc et al. (1996), who found the Bayesian approach to be about twice as fast for their particular specification. Some of the comparisons that we make for mixed logit, such as correlated versus uncorrelated random coefficients and the inclusion of fixed as well as random coefficients, are applicable to probits. Other comparisons, such as the use of lognormal and triangular distributions, are not, since probit and the classical and Bayesian procedures to estimate it depend on the assumption that all random terms are normal.

From both perspectives (though for different reasons) the Bayesian procedure has a theoretical advantage over the classical procedure, independent of convenience. The classical perspective focuses on the sampling distribution of an estimator. As described below, the classical and Bayesian estimators both require integration, though of a different integrand. If the integrals could be evaluated exactly, then the Bayesian and

classical estimators would be on an equal footing from a classical perspective: their sampling distributions are the same asymptotically and cannot be compared theoretically for small samples. However, the integrals cannot be expressed in closed form, and so their evaluation requires simulation, which changes the properties of each estimator. As we discuss below, the conditions under which the simulation-based versions of the estimators become consistent, asymptotically normal and efficient are less stringent for the Bayesian estimator than the classical one. In this regard, the Bayesian procedure has an advantage from a classical perspective. The classical procedure would be preferred only for specifications for which its relative convenience is sufficient to overcome this advantage.

From a Bayesian perspective, the posterior distribution is the relevant concern. The Bayesian procedure provides exact information about the posterior while the classical procedure only approximates the posterior. Though the approximation error disappears asymptotically, much of the appeal of a Bayesian perspective derives from its ability to make statements from small samples as well as large. The classical procedure would only be chosen if it were sufficiently more convenient to warrant the loss from approximation.

In the next section, we specify a model that is about equally easy to handle with either method. We describe the Bayesian and classical procedures for this model, apply them, and present results. In the subsequent sections, we modify the specification in several ways and show how each procedure is affected by the modification. These changes generalize the Bayesian procedure of Sawtooth Software (1999). Run-time comparisons and a discussion of programming changes elucidate the issues associated with applying each method.

2 Mixed Logit with Independent Normal Coefficients

Person n faces a choice among J alternatives in each of T time periods. The person's utility from alternative j in period t is

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt}, \tag{1}$$

where $\varepsilon_{nit} \sim$ iid extreme value and $\beta_n \sim N(b, \Omega)$. The vectors of variables x_{njt} and coefficients β_n have length K. Person n chooses alternative i in period t if $U_{nit} > U_{njt} \forall j \neq i$. Denote the person's chosen alternative in period t as y_{nt}, the person's sequence of choices over the T time periods as $y_n = \langle y_{n1}, \ldots, y_{nT} \rangle$, and the set of $y_n \forall n$ as Y. Conditional on β_n, the probability of person n's sequence of choices is the product of standard logit formulas:

$$L(y_n \mid \beta_n) = \prod_t \frac{e^{\beta_n' x_{ny_{nt}t}}}{\sum_j e^{\beta_n' x_{njt}}}. \tag{2}$$

The unconditional probability is the integral of $L(y_n \mid \beta_n)$ over all values of β_n weighted by the density of β_n:

$$P_n(y_n \mid b, \Omega) = \int L(y_n \mid \beta_n) g(\beta_n \mid b, \Omega) d\Omega. \tag{3}$$

where $g(\cdot)$ is the multivariate normal density. We assume in this section that Ω is diagonal.

2.1 Classical procedure

The log-likelihood function in b and Ω is

$$LL(b,\Omega) = \sum_n \ln P_n(y_n \mid b,\Omega). \tag{4}$$

The integral in $P_n(y_n \mid b,\Omega)$ is approximated through simulation, by taking draws from $g(\cdot)$, calculating $L(y_n \mid \beta_n)$ for each draw, and averaging the results. The simulated probability is designated $\tilde{P}$, and the simulated log-likelihood is

$$SLL(b,\Omega) = \sum_n \ln \tilde{P}_n(y_n \mid b,\Omega). \tag{5}$$

The maximum simulated likelihood estimator (MSLE) is the value of b and Ω, denoted $\hat{b}$ and $\hat{\Omega}$, that maximizes *SLL*. MSLE is consistent if the number of draws used in simulating $P_n(y_n \mid b,\Omega)$ rises with N and is asymptotically normal and efficient, equivalent to the maximum likelihood estimator (MLE), if the number of draws rises faster than $\sqrt{N}$ (Hajivassiliou and Ruud, 1994; McFadden and Train, 2000). If the researcher wants information about β_n for the sampled people, the procedure described by Revelt and Train (2000) can be used.

2.2 Bayesian procedure

The prior on b and Ω is specified. For our purposes, we assign a flat prior on b (either a improper uniform prior, a proper uniform prior over a sufficiently large region, or a proper normal prior with sufficiently large variance that it is effectively flat from a numerical perspective) and assume that the prior on each element of diagonal Ω is inverted gamma with 1 degrees of freedom and scale parameter 1. Let $IG(M \mid m_0, M_0)$ denote the joint density of the diagonal elements of matrix M, each of which is independently distributed inverted gamma with common m_0 degrees of freedom and scale parameter equal to the corresponding element of vector M_0. (This belabored notation facilities generalization to correlated coefficients in the next section.) With these priors, the joint posterior on $\beta_n \forall n$, b and Ω is

$$\Lambda(\beta_n \forall n, b, \Omega \mid Y) \propto \prod_n (L(y_n \mid \beta_n) g(\beta \mid b,\Omega) IG(\Omega \mid 1, \ell) \tag{6}$$

where ℓ is a K-dimensional vector of ones.

Information about the posterior is obtained through simulation, that is, by taking draws from the posterior and calculating relevant statistics, such as moments, over these draws. Gibbs sampling is used to facilitate the taking of draws. In particular, draws are taken sequentially from the conditional posterior of each parameter given the previous draw of the other parameters. The sequence of draws from the conditional posteriors converges to draws from the joint posterior.

The conditional posterior distributions in our model are especially convenient. Given β and Ω, the posterior on b is $N(\bar{\beta}, \Omega/N)$ with $\bar{\beta} = (1/N)\sum\beta_n$, by the standard Bayesian theory of normals (e.g., Zellner, 1971, pp. 14-15 and 20.) This distribution is easy to draw from: A draw of the k-th element of b is created as $\tilde{b}_k = \bar{\beta}_k + (\omega_k/\sqrt{N})\eta$, where η is a draw from a standard normal density and ω_k^2 is the k-th diagonal element of Ω. A draw of the vector b requires only K draws from a random number generator, K means over N terms each, and a few arithmetic calculations. It takes a tiny fraction of a second.

Given b and β, the conditional posterior of Ω is $IG(\Omega \mid 1+N, \ell + N\bar{V})$, where $\bar{V} = (1/N)\sum(\beta_n - b)^2$ is the variance of the β_n's around the given b rather than their own mean β. (See Zellner, 1971, pp. 22-23.) This posterior is also easy to draw from. For each diagonal element of Ω, take $1+N$ draws of iid standard normal deviates, labeled $\eta_r, r = 1, \ldots, (1+N)$, and create $\tilde{\omega}_k^2 = (1+N\bar{V}_k)/\sum_r(\eta_r)^2$. This calculation is also extremely fast.

The only computationally intensive part is drawing $\beta_n \forall n$. Given b and Ω, the conditional posterior for β_n is proportional to $L(y_n \mid \beta_n)g(\beta_n \mid b, \Omega)$. The Metropolis-Hasting (M-H) algorithm is used to take draws from this distribution. (See Chib and Greenberg, 1995, for a general explanation of the M-H algorithm.) The previous draw is labeled β_n^0 and the new one is β_n^1. The new draw is obtained as follows.

1. Calculate $d = \sigma L\eta$, where η is a draw of a K-dimensional vector of iid standard normal deviates, L is the diagonal matrix of square roots of Ω, and σ is a scalar that the researcher sets in a way to be described below.

2. Create a "trial" value of β_n^1 as $\tilde{\beta}_n^1 = \beta_n^0 + d$.

3. Evaluate the posterior at this trial value and compare it with the posterior at the previous draw. That is, calculate the ratio

$$R = \frac{L(y_n \mid \tilde{\beta}_n^1)g(\tilde{\beta}_n^1 \mid b, \Omega)}{L(y_n \mid \beta_n^0)g(\beta_n^0 \mid b, \Omega)}.$$

4. Take a draw from a standard uniform and label the draw μ.

5. If $\mu < R$, accept the trial draw. Otherwise, reject the trial draw and use the previous draw as the current draw. That is, set $\beta_n^1 = \tilde{\beta}_n^1$ if $\mu < R$ and set $\beta_n^1 = \beta_n^0$ otherwise.

A sequence of draws taken by the M-H algorithm converges to draws from the target distribution, in this case the conditional posterior. One draw of β_n within the M-H algorithm for each person is taken in each iteration of the Gibbs sampling over b, Ω, and $\beta_n \forall n$. Movement to convergence in the M-H algorithm for each person and in the overall Gibbs sampling is thereby attained simultaneously.

The value of σ in step (1) affects the acceptance rate in the M-H algorithm. For smaller values of σ, the acceptance rate is generally higher but the jumps between draws is smaller so that more draws are needed for the algorithm to reach convergence and, once at convergence, to traverse the conditional posterior. Gelman et al. (1995) found that the optimal acceptance rate is .4 for $K = 1$ and decreases to .23 for higher dimensions. They recommend an adaptive acceptance rate to achieve optimality. This adaptation is implemented by changing σ in each iteration of the Gibbs sampling based on the acceptance rate among the N trial draws of $\beta_n \forall n$ in the previous iteration. Fol-

lowing Sawtooth Software (1999), we lower σ if the acceptance rate is below .3 and raise it if the rate is above .3.

2.3 Posterior Mean as a Classical Estimator

The Bayesian procedure provides draws from the joint posterior of the parameters. In a Bayesian analysis, these draws are used in a variety of ways depending on the purpose of the analysis. The mean and standard deviation of the draws are simulated approximations to the mean and standard deviation of the posterior. These statistics have particular importance from a classical perspective, due to the Bernstein-von Mises theorem. Consider a model with parameters θ whose true value is θ^*. The maximum of the likelihood function is $\hat{\theta}$, and the mean of the posterior is $\bar{\theta}$ for a prior that is proper and strictly positive in a neighborhood of θ^*. Three interrelated statements are established in different versions of the theorem (e.g., Rao, 1987; Le Cam and Yang, 1990; Lehmann and Casella, 1998; Bickel and Doksum, 2000.)

1. The posterior distribution of θ converges to a normal distribution with covariance B^{-1}/N around its mean, where B is the information matrix. Stated more precisely: $\sqrt{N}(\theta - \bar{\theta}) \xrightarrow{d} N(0, B^{-1})$, where the distribution that is converging is the posterior rather than the sampling distribution.

2. The posterior mean converges to the maximum of the likelihood function: $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$. This result is a natural implication of statement (1). Asymptotically, the shape of the posterior becomes arbitrarily close to the shape of the likelihood function, since the posterior is proportional to the likelihood function times the prior and the prior becomes irrelevant for large enough N. The mean and mode of a normal distribution are the same.

3. The asymptotic sampling distribution of the posterior mean is the same as for the maximum of the likelihood function: $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, B^{-1})$. This result is obvious from statement (2).

The third statement says that the mean of the posterior is an estimator that, in classical terms, is equivalent to MLE. The first statement establishes that the standard deviations of the posterior provide classical standard errors for the estimator.

The true mean and standard deviation of the posterior cannot be calculated exactly except in very simple cases. These moments are approximated through simulation, by taking draws from the posterior and calculating the mean and standard deviation of the draws. For fixed number of draws, the simulated mean, denoted $\breve{\theta}$, is consistent and asymptotically normal, with variance equal to $1 + (1/R)$ times the variance of the non-simulated mean, where R is the number of (independent) draws. If the number of draws (whether independent or not) is considered to rise with N at any rate, the simulation noise disappears asymptotically such that $\breve{\theta}$ is efficient and asympototically equivalent to MLE. In contrast, MSLE is inconsistent for a fixed number of draws. For consistency, the number of draws must be considered to rise with N, but even this condition is not sufficient for asymptotic normality. The number of draws must be considered to rise faster than $\sqrt{N}$ for MSLE to be asymptotically normal, in which case it is also equivalent to MLE. Since it is difficult to know in practice how to satisfy the condition that the number of draws rises faster than $\sqrt{N}$, $\breve{\theta}$ is attractive relative to

MSLE, even though their non-simulated counterparts are equivalent.

2.4 Application

We apply both estimators to data on customers' choice among energy suppliers. The data were collected by RTI (1997) for the Electric Power Research Institute. Each of 361 customers were 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/she 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.

Table 1 gives the MSLE and the simulated mean of the posterior, with standard errors for each. Recall that each coefficient is assumed to be independently normal, such that the parameters are the mean and standard deviation in the population of each random coefficient. The two procedures provide similar results. The scale of the estimates from the Bayesian procedure is somewhat larger than that for MSLE. 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 point estimates as well as standard errors. Run time was essentially the same for each approach. In fact, independent normals were specified in this section because the two methods under this specification took about the same run time and were about equally easy to program. With other specifications, run times and programming effort differed, as described below. However, the results were similar from the two methods under all specifications.

3 Multivariate Normal Coefficients

The behavioral model is the same except that Ω is full rather than diagonal. The classical procedure is the same except that drawing from $g(\beta_n \mid b, \Omega)$ for the simulation of the probability $P(y_n \mid b, \Omega)$ requires creating correlation among independent draws from a random number generator. The model is parameterized in terms of the Choleski factor of Ω, labeled L where $LL' = \Omega$, since doing so assures that the resulting Ω is always positive definite. 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, the main difference is that the model has far more parameters: $K + K(K+1)/2$

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 Table 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 Ω, the inverted gamma distribution is replaced with its multivariate generalization, the inverted Wishart. We specify the prior as inverted Wishart with K degrees of freedom and parameter KI where I is the identity matrix. This density is denoted $IW(\Omega \mid K, KI)$. The posterior is then $IW(\Omega \mid K+N, KI+N\bar{V})$, where $\bar{V} = (1/N)\sum(\beta_n - b)(\beta_n - b)'$. Draws from the inverted Wishart are easily obtained. Take $K+N$ draws of K-dimensional vectors of iid standard normal deviates. Calculate the Choleski factor, M, of $(KI+N\bar{V})^{-1}$. Create $S = \sum_r (M\eta_r)(M\eta_r)'$. Then $\tilde{\Omega} = S^{-1}$ is a draw. Note that draws from an inverted gamma are obtained by a one-dimensional version of this process. The only extra computer time relative to independent coefficients arises in the calculation of a K-by-K moment matrix for $\bar{V}$ and a Choleski factor rather than K variances. This difference is trivial for typical values of K. As shown in Table 2, run time for the model with full covariance among the random coefficients was essentially the same as with independent coefficients.

4 Fixed Coefficients for Some Variables

There are various reasons that an analyst 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 the economically meaningful concept. He recommends holding at least one coefficient fixed. (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 log-normals, 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 Ω 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 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 the fixed coefficients 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

$$U_{njt} = \alpha' z_{njt} + \beta_n' x_{njt} + \varepsilon_{njt}, \tag{7}$$

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

$$L(y_n \mid \alpha, \beta_n) = \prod_t \frac{e^{\alpha' z_{ny_{nt}t} + \beta_n' x_{ny_{nt}t}}}{\sum_j e^{\alpha' z_{njt} + \beta_n' x_{njt}}}. \tag{8}$$

The conditional posteriors for Gibbs sampling are:

(i) $\Lambda(\beta_n \mid \alpha, b, \Omega) \forall n \propto L(y_n \mid \alpha\beta_n) g(\beta_n \mid b, \Omega)$. 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 in the calculation of the relative density R.

(ii) $\Lambda(b \mid \beta_n, \Omega) = N(0, \Omega/N)$. Note that α does not enter this posterior; its effect is incorporated into the draws of β_n from layer (i).

(iii) $\Lambda(\Omega \mid \beta_n, b) = IW(K+N, KI+N\bar{V})$. Again, α does not enter directly.

(iv) $\Lambda(\alpha \mid \beta_n) \propto \prod_n L(y_n \mid \alpha, \beta_n)$. Draws are obtained with M-H on the pooled data.

Layer (iv) takes as much time as layer (i), since each involves calculation of a logit formula for each observation. H-B 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 2, this expectation is confirmed in our application.

5 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

$$U_{njt} = (e^{\beta_n})' x_{njt} + \varepsilon_{njt}, \tag{9}$$

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

$$L(y_n \mid \alpha, \beta_n) = \prod_t \frac{e^{(e^{\beta_n})' x_{ny_{nt}t}}}{\sum_j e^{(e^{\beta_n})' x_{njt}}}. \tag{10}$$

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 interations 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, and we have not observed any other problems arising. The Gibbs sampling seems to converge a bit more slowly, but not appreciably so. As indicated in Table 2, run time for the classical approach rose nearly 50 percent with lognormal relative to normals (due to more iterations being needed), while the H-B 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.

6 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 is the same as with draws from a normal. Our 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 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 indendent triangular distributions for all coefficients with mean and spread vectors b and s, and flat priors on each, the conditional posteriors are:

i. $\Lambda(\beta_n \mid b,s) \propto L(y_n \mid \beta_n)h(\beta_n \mid b,s)$ 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.

ii. $\Lambda(b,s \mid \beta_n) \propto \prod_n h(\beta \mid b,s)$. 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 layer (i), draws of β_n that are outside the range $(b-s,b+s)$ from layer (ii) are necessarily rejected. And in layer (ii), draws of b and s that create a range $(b-s,b+s)$ that does not cover all the β_n's from layer (i) 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 (2,4) in one iteration of layer (ii), then the next iteration of (i) 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 interations were needed to achieve a semblence of convergence, compared with normal distributions. Run time rose by a factor of four as a result.

7 Conclusions

The Bayesian approach has theoretical advantages from both a classical and Bayesian perspective. 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 also seems to be faster 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.

References

Albert, J., and S. Chib, 1993, "Bayesian Analysis of Binary and Polychotomous Data," *Journal of the American Statistical Association*, Vol. 88, pp. 669-679.

Allenby, G., and P. Rossi, 1999, "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, Vol. 89, Nos. 1-2, pp. 57-78.

Brownstone, D., and K. Train, 1999, "Forecasting New Product Penetration with Flexible Sustitution Patterns," *Journal of Econometrics*, Vol. 89, Nos. 1-2, pp. 109-129.

Bolduc, D., B. Fortin, and S. Gordon, 1996, "Multinomial Probit Estimation of Spatially Interdependent Choices: An Empirical Comparison of Two New Techniques," working paper, Department of Economics, University of Laval.

Chib, S., and E. Greenberg, 1995, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, Vol. 49, pp. 327-335.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, 1995, *Bayesian Data Analysis*, Suffolk: Chapman and Hall, p. 335.

Geweke, M., 1989, "Efficient Simulation from the Multivariate Normal Distribution Subject to Linear Inequality Constraints and the Evaluation of Constraint Probabilities," working paper, Duke University.

Hajivassiliou, V., 1990, "Smooth Estimation Simulation of Panel Data LDV Models," working paper, Yale University.

Hajivassiliou, V., and P. Ruud, 1994, "Classical Estimation Methods for LDV Models using Simulation," in R. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. IV, New York: Elsevier.

Huber, J., and K. Train, 2001, "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths," *Marketing Letters*, Vol. 12, No. 3, pp. 257-267.

Keane, M., 1990, "A Computationally Efficient Practical Simulation Estimator for Panel Data, with Applications to Estimating Temporal Dependence in Employment and Wages," working paper, University of Minnesota.

McCulloch, R., and P. Rossi, 1994, "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, Vol. 64, pp. 207-240.

McFadden, D., and K. Train, 2000, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, Vol. 15, No. 5, pp. 447-470.

RTI, 1997, *Predicting Retail Customer Choices Among Electricity Pricing Alternatives*, Electric Power Research Report, Palo Alto.

Revelt, D., and K. Train, 1998, "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *The Review of Economics and Statistics*, Vol. LXXX, No. 4, pp. 647-657.

Revelt, D., and K. Train, 2000, "Customer-Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier," Working Paper No. E00-274, Department of Economics, University of California, Berkeley.

Ruud, P., 1996, "Approximation and Simulation of the Multinomial Probit Model: An Analysis of Covariance Matrix Estimation," working paper, Department of Economics, University of California, Berkeley.

Sawtooth Software, 1999, "The CBC/HB Module for Hierarchical Bayes Estimation," at *http://www.sawtoothsoftware.com/Techabs.htm*.

Zellner, A., 1971, *Introduction to Bayesian Inference in Econometrics*, New York: Wiley.

Table 1: Mixed Logit Model of Choice Among Energy Suppliers

Estimates (se's in parens.)		MSL	HB	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)

Table 2: Run-Times in minutes

Specification	MSL	HB
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