Bayesian Analysis of the Multinomial Probit Model

by

Robert E. McCulloch and Peter E. Rossi

Graduate School of Business, University of Chicago

April 1996, revised August 1996

Abstract

We review and extend our earlier work on Bayesian analysis of the Multinomial Probit model. We employ a hierarchical Bayesian framework and define various Gibbs samplers for analysis of these hierarchical models. We outline two basic samplers for the MNP model – one which navigates in the full parameter space and one which navigates only in the space of identified parameters. We consider extensions of the basic model to include panel data settings with random effects priors as well as general priors for the covariance matrix of the probit equation errors which offer more flexibility than the standard Wishart Prior.

1. Introduction

In this chapter, we discuss Bayesian Analysis of the Multinomial Probit Model (MNP) using Markov Chain Monte Carlo methods. Although the MNP model has been in the econometrics and psychology literature for some 60 years, it is only recently that estimation and inference methods have made it feasible to analyze MNP models with more than two or three response categories. Classical sampling theoretic approaches to estimation of the MNP model have recently been proposed in the econometrics literature (see Hajivassiliou [1994] for an excellent overview of these methods). All of these classical econometric methods rely on asymptotic approximations to conduct inference about the probit model parameters. McCulloch and Rossi (1994) show that is possible to conduct exact, likelihood-based inference for the MNP model by using a Bayesian simulation method which complements the work by Albert and Chib (1993) on the binomial probit model (see Zellner and Rossi [1984] for a non-simulation approach to Bayesian inference in the binomial setting). Evidence in McCulloch and Rossi (1994) and Geweke, Keane and Runkle (1994) shows that the asymptotic approximations used in the classical approaches can be inaccurate and that the improved inference available in the Bayesian approach is no more computationally demanding than the classical simulation-based approaches.

Our Bayesian method can easily be extended to handle hierarchical or random coefficient models used with panel data, autocorrelated latent regression errors, and non-normal random coefficient distributions all within the same hierarchical framework. Below we outline approaches for each of these important extensions. All of these extensions build on the basic hierarchical model structure laid out in McCulloch and Rossi. A critical component of the Bayesian approach to inference in the MNP is a prior on the covariance matrix of the latent regression errors. We review two basic approaches to specifying these covariance matrix priors. We also review the approach of the Barnard, McCulloch and Meng (1996) and show how this very flexible class of covariance matrix priors can be used in this situation.

2. The MNP Model and Identification

2.1 The Model

To begin, we briefly review the notation for the MNP model. Let Y be a multinomial random variable which takes on one of the possible values, {1, …, p}. In many econometric applications, Y denotes the choice made by economic agents between p alternative goods. We observe Y conditional on information contained in the matrix Z $(p \times k')$. The conditional distribution of Y | Z is specified via a latent regression system.

i. $U = Z\delta + v \quad \text{where } v \sim N(0, \Omega)$

We do not observe U directly but instead observe the index of the maximum W

ii. $Y = i \quad \text{if } \max_i(U) = U_i$

Here max(U) means the maximal element of $U' = \left(U_1, U_2, \ldots, U_p\right)'$.

U is often interpreted in the choice context as the system of latent utilities associated with p choice alternatives. Agents choose among the mutually exclusive p choice alternatives so as to maximize utility but the econometrician is unable to observe the utility levels and must make inferences about utility using choice information alone.

i) and ii) define the sampling model, Y | Z, δ Σ. Typically, we observe the set of observations (Y_i, Z_i) and assume that they are iid according to the sampling model. In many applications, choice data is obtained by observing a panel of consumers and the iid assumption requires modification. In section 7 below, we elaborate the model to accommodate a panel structure.

At this point, we can readily appreciate two problems associated with inference in this model. First, the model as specified in (1) and (2) is not identified. Second, the likelihood of $\delta, \Sigma \mid Y, Z$ requires the evaluation of the conditional multinomial choice probabilities which require integration of a p dimensional normal distribution over sets of the form $\left\{U \middle| Y(U) = i\right\}$. These sets are cones over which the normal distribution is difficult to integrate. Much of the research on the MNP model has been devoted to the development of methods for computing these integrals which would allow for fast evaluation of the likelihood function or moment conditions.

2.2 Identification

As is well-known, the model specified by i) and ii) is not identified. The distribution of $Y|Z$ is unchanged adding a scalar random variable to all components of U or by scale shifts. The location invariance problem is solved by differencing the system which respect to some base choice alternative (which we assume is alternative 1).

$$W_i = U_{i+1} - U_1 \qquad i = 1, \ldots, p-1$$

$$w_i^{'} = z_{i+1}^{'} - z_1^{'}$$

If there are intercept terms in the Z matrix for each choice alternative, then the differencing of the rows of Z expresses each of the intercepts as differences with respect to the first choice alternative. Thus, to achieve identification we customarily set the first element of δ to zero. From this point on, we will call this vector, β. The sampling model can now be written as

$$W = X\beta + \varepsilon \qquad \varepsilon \sim N(0,\Sigma)$$
$$Y = \begin{cases} 0 & \text{if } \max(W) < 0 \\ i & \text{if } \max(W) = W_i > 0 \end{cases}$$

W is p-1, X $((p-1)\times k)$, and Σ $((p-1)\times(p-1))$.

While differencing the system removes the location invariance problem, we still face the problem of scale invariance. This problem arise from the fact that Y(cW) = Y(W) for all c > 0. Since $cW = c(X\beta+\varepsilon) = X(c\beta) + c\varepsilon$, we can see that $Y \mid X, \beta, \Sigma = Y \mid X, c\beta, c^2\Sigma$ and thus that $L(\beta,\Sigma) = L(c\beta,c^2\Sigma)$.

In the classical literature, it is common to fix an element of Σ (e.g set $\sigma_{11} = 1.0$) (c.f. Dansie[1985]). Another possibility would be to fix an element of the β vector which would require a priori knowledge of the sign of that element. In the Bayesian approach, it is not straightforward to adopt a prior with $\sigma_{11} = 1.0$ since this requires putting a prior on the set of covariance matrices with fixed 1,1 element. As we show below, this is possible but it requires non-standard, non-conjugate priors.

3. The MCMC Approach to Bayesian Inference in the MNP

To conduct Bayesian inference in the MNP, we must summarize information given by the posterior distribution of the model parameters.

$$p(\beta,\Sigma|Y,X) \propto p(\beta,\Sigma)p(Y|X,\beta,\Sigma)$$

Generally, we want to compute the moments of the posterior and make various probability statements about the likely range of parameter values. If the likelihood function could be evaluated at low computational cost, we could use standard numerical integration methods as in Zellner and Rossi (1984). Even with more modern methods of approximating the probabilities need to evaluate the likelihood function, direct numerical integration procedures would be computationally infeasbile for all except the smallest models. In addition, approximation error in computation of the multinomial probabilities would have to be held to a minimum to as not to affect the properties of the direct numerical integration methods such as importance sampling.

3.1 Gibbs Samplers for the MNP

Our approach to the problem of posterior inference for the MNP model is to construct a method which provides the equivalent of a indirect simulator from the posterior distribution. We construct a Markov chain whose invariant or equilibrium distribution is the posterior distribution. Thus, we can simply run this Markov chain forward from some starting point to generate sequences of draws which can be used to estimate any desired feature of the posterior distribution. This idea of constructing a Markov chain as a indirect posterior simulator is called Markov Chain Monte Carlo (MCMC) and has been fruitfully applied to many important problems in Bayesian statistics and econometrics.

In the analysis of the MNP model, we use a particular MCMC method called the Gibbs sampler (introduced by Geman and Geman[1984] and Tanner and Wong [1987]). The Gibbs sampler relies on the remarkable result that iterative, recursive sampling from the full set of conditional distribution results in a Markov chain with equilibrium distribution equal to the joint distribution. In the case of the MNP model, if we *augment* the parameters (β,Σ) with the vector of all latent utilities, W, then we can break the full set of parameters into three groups and draw these groups successively to form the Gibbs sampler. This strategy relies on the fact that, conditional on W, the Bayesian analysis of the MNP reduces to standard linear model results. The three groups of conditional distributions are defined as follows:

i. $w \big| \beta, G = \Sigma^{-1}, y, X$ where $w' = \left(W_1', \ldots, W_N'\right), y' = \left(Y_1', \ldots, Y_N'\right),$

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$$

ii. $\beta \big| w, X, G$

iii. $G \big| \beta, W$

To implement the Gibbs sampler, we start with values of β and G. We then draw w from i, β from ii using the new value of w, and finally G from iii using the new values of w and β. This process is repeated to produce a sequence of draws of (w, β, G). Note that since this simple version of the MNP model is iid, we can draw the w vector in conditional i) above, one observation at a time and then piece the W_i together.

The distribution of $W_i \big| \beta, G, Y_i$ is a p-1 dimensional normal distribution truncated to a cone. For example, if Y_i =j then $W_{i,j} > \max\left(W_{i,-j}, 0\right)$ where $W_{i,-j}$ is the p-2 dimensional vector of all of the components of W excluding $W_{i,j}$. We avoid this problem of drawing from a truncated multivariate normal distribution by using a sub-chain Gibbs sampler to draw from the truncated univariate conditional distributions, $W_{i,j} \big| W_{i,-j}, \beta, \Sigma, Y_i, X_i$. The truncation point of this univariate normal distribution can readily be calculated from the definition of implied regions in $\Re^{p-1}$. The moments of these univariate normal distributions can readily be calculated from the multivariate normal distribution of $W_i \big| \beta, G, X_i \sim N\left(X_i\beta, G\right)$. To make draws from the truncated univariate normal distribution, we use a rejection strategy (see McCulloch and Rossi [1994], p. 212).

The second conditional distribution, $\beta | G, w$, is a simple matter to draw from if the prior on β is multivariate normal. $\beta \sim N\left(\overline{\beta}, A^{-1}\right)$.

Let G = C C'

$$\beta \Big| w, X, G \sim N\left(\hat{\beta}, \Sigma_\beta\right) \qquad \Sigma_\beta = \left(X_*' X_* + A\right)^{-1}; \hat{\beta} = \Sigma_\beta\left(X_*' w_* + A\overline{\beta}\right)$$

$$w_* = \iota_N \otimes C' w;\ X_* = \iota_N \otimes C' X$$

The method by which we draw from the conditional posterior distribution of G depends critically on the prior adopted. In this chapter, we develop three different sorts of priors for G, each with a different draw strategy which we introduce in sections 4, 5, and 6 below.

To summarize, we use various Gibbs samplers to construct a Markov chain which enables us to estimate via simulation any required posterior quantity. Given sufficient computer resources, we can achieve a very high degree of accuracy in these estimates of posterior quantities. Thus, while our inferences do not rely on asymptotics in the sense of arbitrarily large sample sizes, we do require large simulation sizes to achieve a high degree of accuracy in the estimates of posterior quantities. We take the basic view that asymptotic approximations are more relevant and useful when the investigator has control over the sample size.

3.2 Practical Considerations for the MNP Gibbs Sampler

The theory of MCMC and, in particular, the Gibbs sampler shows that under very mild conditions, the Gibbs sampler Markov Chain will converge in distribution to the posterior distribution at a geometric rate (see McCulloch and Rossi [1984] and Tierney [1991]). While these theoretical results assure the eventual convergence of the MNP Gibbs samplers, this theory offers little practical guidance. There are two important practical considerations: 1) how long must the Gibbs sampler be run in order to be confident that the effect of initial conditions have dissipated? 2) what is the information content of a given sequence of Gibbs draws from the stationary distribution?

These practical convergence considerations can only be answered empirically. To assess the rate at which the initial conditions are dissipated, we conduct an analysis of the sensitivity of the estimated posterior distributions to various, widely dispersed initial conditions (see Gelman and Rubin [1992]). For each initial condition, we "burn-in" or run out the Gibbs sampler for a large number of draws (typically at least 5000 draws) and then use the remaining draws in the sequence to estimate the posterior distribution. If we have chosen an adequate burn-in series length, T*, then the estimated posteriors should all be about the same for all starting points. McCulloch and Rossi [1994] report a series of experiments in which the algorithm specified in section 3 below is tested with a wide variety of initial conditions. The sampler appears to converge rapidly from any initial condition. Our experience with more complicated samplers affirms this general finding of insensitivity to initial conditions (Nobile[1995] documents a possible exception which can occur using informative priors, see section 4 for details).

If we feel comfortable that the "burn-in" periods of T* draws is sufficient, we then use the remaining T-T* draws to estimate posterior quantities. It is important to remember that the Gibbs sampler is a non-iid simulation method and, therefore, the draws can exhibit dependence. In the first applications of the Gibbs sampler to linear models, this dependence in the draw sequence was never a severe problem and quite short runs could be made. However, our experience with the MNP model is that the draw sequences can be highly autocorrelated, necessitating long runs of the sampler to achieve good accuracy in the estimates of posterior quantities. Given even relatively modest computing resources, it is possible to make very long runs of the MNP Gibbs samplers even for relatively large samples and high dimensional (e.g. p>4 and N > 2000) problems. Our usual approach is to make a "burn-in" run and then start short runs on a number of workstations, starting from the "burned-in" values of the parameters (including the latent variables). These short runs are pieced together to form the draw sequence which is used for inference. In this manner, we can solve most problems in no more than one day of computing.

4. An Algorithm with Non-identified Parameters

As discussed in section 2.2, the model parameters (β,Σ) are not identified. The identified parameters are functions of these unidentified parameters.

$$\tilde{\beta}=\beta/\sqrt{\sigma_{11}};\ \tilde{\Sigma}=\Sigma/\sigma_{11}$$

If we desire informative priors, it would seem most convenient to put these priors directly on the identified parameters, $(\tilde{\beta},\tilde{\Sigma})$. This requires a prior on the space of covariance matrices conditional on σ_{11} =1.0. One approach to this problem taken in McCulloch and Rossi (1994) is to specify a prior on the full set of parameters (β,Σ). This induces a prior on the identified parameters. For a given choice of the prior hyperparameters, we then examine the implied prior on the identified parameters. With some trial and error, we can hope to find prior parameter settings that will appropriately reflect our views on the identified parameters. In practice, this approach is most useful for specifying relatively diffuse or uniformative priors.

To define the Gibbs sampler in this case, we use the approach outlined in 3.1 and specify a conditionally conjugate Wishart prior on G.

$$p\left(G\middle|\nu,V\right)\propto\left|G\right|^{\frac{\nu-p-1}{2}}\operatorname{etr}\left\{-\frac{1}{2}GV\right\}$$

The posterior of G is also in the Wishart form.

$$G\left|\beta,w,X\sim W\left(\nu+N,V+\sum_{i=1}^{N}\varepsilon_i\varepsilon_i'\right)\right.$$

If we combine this Wishart draw with the draws from the conditional posterior of β and the draws of w specified in section 3.1, we define a Gibbs sampler that navigates in the full parameter space of unidentified parameters. Fortran code for this algorithm can be obtained at ftp://gsbper.uchicago.edu in the directory pub/rossi/mnp.

Our approach in most applications of this sampler is to specify proper but fairly diffuse priors on G and β. We investigate the induced prior over the identified parameters to insure that it reflects a proper level of diffusion. We then run the Gibbs sampler to indirectly simulate from the joint posterior of (β,Σ,w). We report the *marginal* posterior distribution of the identified parameters. It should be emphasized that it does not matter whether we first marginalize the prior and conduct the analysis only on the identified parameters or marginalize the posterior since the likelihood depends only on the identified parameters.

This Gibbs sampler navigates freely in the non-identified parameter space, constrained only the non-identifed directions only by the proper prior. Nobile (1995) has constructed some examples in which this algorithm can take a very long time to dissipate the initial conditions. These examples involve informative priors on the space of non-identified priors. These priors "flatten" down the likelihood ridge in certain regions of the parameter space and make it difficult for the algorithm to find the region of high posterior value. The priors Nobile considers are informative about non-identified parameters. It is hard to imagine how such prior information could arise in normal circumstances. Nobile proposes a hybrid MCMC method which accelerates convergence.

5. An Algorithm with Fully Identified Parameters

An alternative to the non-identified algorithm would be to specify a prior directly on the identified parameters, $(\tilde{\beta},\tilde{\Sigma})$, in such as way as we can easily draw from the appropriate conditional posteriors. McCulloch, Polson and Rossi (1993) introduce a computationally attractive prior for $\tilde{\Sigma}$.

We define this prior by first reparameterizing Σ. According to the basic model, $\varepsilon_i \sim N(0,\Sigma)$. Write $\varepsilon_i' = \{\varepsilon_{i,1},(\varepsilon_{i,2},\cdots,\varepsilon_{i,p-1})\} = \{U_i, e_i'\}$. We can then write the joint distribution of ε_i as the marginal distribution of U_i and the conditional distribution, $e_i|U_i$. Let $\gamma = E[U_i e_i]$ and $\Sigma_e = E[e_i e_i']$. We then can write the marginal distribution of U and the conditional distribution of $e_i|U_i$.

$$U_i \sim N(0,\sigma_{11})$$

$$e_i|U_i \sim N\left(\frac{\gamma}{\sigma_{11}}U_i, \Sigma_z - \gamma\gamma'/\sigma_{11}\right)$$

Let $\Phi = \Sigma_z - \gamma\gamma'/\sigma_{11}$. Then we can reparameterize Σ in terms of σ_{11}, γ, and Φ.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \gamma' \\ \gamma & \Phi + \gamma\gamma'/\sigma_{11} \end{bmatrix}$$

Hence, we can put a prior on $\{\Sigma|\sigma_{11}=1\}$ by setting $\sigma_{11}=1$ and putting priors on γ and Φ.

For convenience, we use the following priors

$$\gamma \sim N(\bar{\gamma}, B^{-1})$$

and

$$\Phi^{-1} \sim W(\kappa, C)$$

To obtain draws from $(\gamma, \Phi)|\beta, w, X$, we note that we "observe" (U_i, e_i). γ and Φ are simply the parameters of a multivariate regression of e_i on U_i. Standard results from Bayes linear models provide the conditional posteriors -

$$\gamma|\Phi, \beta, w, X \sim N\left(A_\gamma\left(\left(\mathrm{vec}(\Phi^{-1}E'U\right) + B\gamma\right), A_\gamma\right)$$

where $A_\gamma = \left(U'U\Phi^{-1} + B\right)^{-1}$; $U' = (U_1, \ldots, U_N)$; $E' = (e_1, \ldots e_N)'$

and

$$\Phi^{-1}|\gamma, \beta, w, X \sim W\left(\kappa + N, C + (E - U\gamma')'(E - U\gamma')\right)$$

The Gibbs sampler for the identified parameter case is defined as the following sequence of conditional draws.

i. $w|\tilde{\beta}, \tilde{\Sigma}, X, y$

ii. $\tilde{\beta}|\tilde{\Sigma}, w, X$

iii. $\tilde{\Sigma}|\tilde{\beta}, w, X$ which is achieved via

$$\gamma|\Phi^{-1}, \tilde{\beta}, w, X \quad \text{and} \quad \Phi^{-1}|\gamma, \tilde{\beta}, w, X$$

This Markov chain navigates in a lower (by one) dimensional space that the MNP sampler for non-identified parameters. In addition, the one-shot draw of the covariance matrix is broken down into two conditional draws which may introduce an additional source of autocorrelation into the Markov chain. It should be noted, however, that with the natural conjugate prior of the multivariate regression framework (see Zellner [1971], chapter VIII) we can draw $\tilde{\Sigma}$ in one shot by drawing from the marginal posterior of $\Phi^{-1}\big|\tilde{\beta}, \mathrm{w}, \mathrm{X}$ instead of the posterior conditional on γ.

Some care has to be used in assessing the prior hyperparameters in this set-up. The relative diffusion of the prior on γ and Φ is important. To see this, consider the case of p=3.

$$\tilde{\Sigma} = \begin{bmatrix} 1 & \gamma \\ \gamma & \phi + \gamma^2 \end{bmatrix}$$

and

$$\mathrm{corr}(\varepsilon_1, \varepsilon_2) = \frac{\gamma}{\sqrt{\phi + \gamma^2}}$$

If the prior on γ is relatively more diffuse that the prior on ϕ, then the joint prior will put most mass near high correlations.

The advantages of the prior on identified parameters is twofold: 1) improper diffuse priors are possible and 2) informative priors for $\tilde{\beta}$ more easily assessed than for the method which uses the full non-identified parameterization. In many applications, we have prior information on the likely values of $\tilde{\beta}$. In the prior set-up for the non-identified parameters, we do not directly assess a prior on the identified slope coefficients but must induce a complicated prior instead (see McCulloch, Polson, and Rossi [1993] for an analysis of the implied marginal distribution of $\tilde{\beta}$ for this prior).

One might ask the legitimate question of whether or not it is possible to asssess a prior on the identified parameters via the method of section 4 which is similar to the prior introduced in this section. In McCulloch, Polson, and Rossi (1993) the marginal distributions of γ and Φ are derived under the non-identified prior. While the priors on the identified parameters are not identical for the two methods in section 4 and 5, it is possible to assess very similar priors by careful equating of moments.

6. Other Informative Priors for the Covariance Matrix

Experience with actual and simulated data has shown that it is very difficult to make precise inferences about the covariance parameters in the MNP. Keane (1990) notes that there are situations in which these covariance parameters are nearly unidentified. Our experience is that the relative variances ($\sigma_{jj} / \sigma_{11}$) are well identified by the data, but that the correlations may require a great deal of information to make relatively precise inferences about. Thus, in many situations, it may be desirable to inject prior information about the correlation patterns in the data via exact restrictions on covariance structure or via informative priors.

6.1 Variance Components Approaches

To discuss the variance component approach, it is useful to return to the notation for the undifferenced system of latent variables.

$$U = Z\delta + v \qquad v \sim N(0, \Omega)$$

If we start with an independence probit (Ω=D, a diagonal matrix), then the differenced system errors will show an equi-correlated structure.

$$\operatorname{Var}(\varepsilon = v_{-1} - v_1) = \begin{bmatrix} d_2 + d_1 & d_1 & \cdots & d_1 \\ & d_3 + d_1 & \ddots & \vdots \\ & & \ddots & d_1 \\ & & & d_{p-1} + d_1 \end{bmatrix}$$

If D = dI, we would have an equi-correlated structure with a correlation of .5.

Patterns of correlation can be most easily introduced by using a variance components structure. For the purpose of interpretation, we find it more intuitive to introduce the variance components in the undifferenced system. We can think of grouping the p choice alternatives into G groups, each of which has some common, unobservable component of utility. For example, if the alternatives are different brands of a consumer product, then the high-priced products might all enjoy a higher perceived quality in the eyes of certain consumers. This would introduce a variance component for quality perceptions for the higher-priced brands.

We break the p alternatives into G groups. Let g(j) be an indicator index function for group g. $g(j) = 1$ if $j \in$ indices for group g, g=1, …, p.

$$v_j = w_j + \sum_{g=1}^{G} g(j) r_g$$

$$w_j \sim \text{iid}\, N\left(0, \sigma_w^2\right); \qquad r_g \sim N\left(0, \sigma_g^2\right); \qquad \{r_g\} \text{ are independent.}$$

This parameterizes the **Ω** matrix with G+1 variance component parameters. We note that because of the restricted structure of this parameterization of **Ω**, differencing is not required for identification purposes. We still must impose some restriction on the covariance parameters for identification. If we restrict $\sigma_W^2 = 1$, this will achieve identification.

We can easily define a Gibbs strategy for this variance component structure by introducing the $\{r_g\}$ as additional latent variables. If we group together the components of each v_i vector corresponding to group g, then the conditional posterior of r_g is a normal mean problem. Our MNP Gibbs sampler for the variance component problem replaces the draw of Σ or $\widetilde{\Sigma}$ with the following sets of conditional draws:

$$r_{g,i} \mid v_i, \sigma_g$$

$$\sigma_g \left| \left\{ r_{g,i} \right\} \right.$$

6.2 More General Priors on the Correlation Matrix

The variance component approach to specifying an information prior on the covariance structure is appealing because of it's parsimony and usefulness in cases in which the alternatives can be grouped into similar groups. However, the covariance structure generated from the variance component model is restrictive in that it only affords positive covariances. In addition, a prior information on grouping may not be available. A more general approach would be to put an informative prior on the matrix of correlations which "shrinks" the correlation structure toward some base case such as the equi-correlated case produced by a scalar independence probit.

Barnard, McCulloch and Meng [1996] introduce a general class of priors for covariance matrices that can be applied fruitfully here. To introduce this prior, it is most convenient to return to the differenced system. The Barnard et al approach is to reparameterize the covariance matrix in terms of the standard deviations and the correlation matrix and then put independent priors on each.

$\Sigma = \Lambda R \Lambda$ where Λ is a diagonal matrix with $\lambda_{ii} = \sigma_{ii}$

The prior considered in Barnard et al is

$$p(\Lambda, R) = p(\Lambda)p(R); \quad p(R) \propto 1; \;\; \log(\lambda_i) \sim \text{iid}\, N(\mu_i, \tau_i)$$

The emphasis in Barnard et al is on the prior for the variances with the primary motivation stemming from a location shrinkage application. Here we might prefer to be very diffuse on the variances and more informative on R.

In this general framework, we put a prior on Σ by simply specifying p(Λ) and p(R). We set $\lambda_1 = 1$ for identification purposes, and let the log-normal priors on remaining elements of Λ be very diffuse (or even improper). We then must choose an appropriate prior on R. The region of support of the prior on R is a subset of the hypercube with side (-1,1). This region becomes more restricted as you move away from the origin due to the constraints implied by positivity of the R matrix. This implies the "non-informative" uniform prior on R is actually informative due to the nature of the restricted support. For p>2, the marginal prior distributions of each r_{ij} element is non-uniform and becomes more concentrated near zero as the dimension (p) increases. If we wanted to "center" the prior over the independent scalar Probit model (this is the probit analogue of the multinomial logit model), we would have to use a prior which puts mass around .5. One such prior with reasonably slow damping tails might be

$$p(R) \propto \exp\left\{-\eta\sum_{i=1}^{p-1}\sum_{j=i+1}^{p-1}\left|r_{ij}-.5\right|\right\}.$$

The task of drawing from the conditional posterior, $\tilde{\Sigma}\left|\tilde{\beta}, w, X\right.$, might seem formidable because of the reparameterization of Σ, the non-conjugate nature of the prior, and the restricted region of support. The un-normalized conditional posterior density can be written.

$$p(\tilde{\Sigma}\left|\tilde{\beta}, w, X\right.) = p(\tilde{\Sigma}\left|E\right.) = p(\Lambda, R\left|E\right.) \propto \operatorname{etr}\left\{-.5(\Lambda R\Lambda)^{-1}EE'\right\}p(R)\prod_{i=2}^{p-1}p(\lambda_i)$$

$$E = \left[\varepsilon_1, \ldots, \varepsilon_N\right]$$

Direct draws from this density are not possible. A very general solution to this problem is to define a sub-Gibbs sampler and draw the elements of Λ and R one by one, conditional on the others. This one-by-one method also allows one to compute the region of support for each r_{ij} given all of the other elements. Still, the conditional distributions are in very non-standard forms. A highly effective solution to this problem is to adopt a "Griddy" Gibbs strategy (Ritter and Tanner, 1992). The Griddy Gibbs sampler uses a discrete approximation to the conditional distributions used in the sub-Gibbs sampler. We choose a grid of points in the support of the prior and simply compute the multinomial approximation to the conditional distribution by evaluating the conditional posterior density at each of the grid points. It should be emphasized that the Griddy Gibbs Sampler does not suffer from the direct curse of dimensionality that would afflict a standard discrete approximation to the joint posterior. We only evaluate the conditional posteriors, one by one, and never have to evaluate the joint posterior at a grid of points designed to fill the entire parameter space. Furthermore, the grids can be adaptive so that we only put points in regions where there is high posterior mass. Barnard et al report effective use of the Griddy Gibbs method for up to 10 x 10 covariance matrices.

7. Bayesian Random Coefficient or Hierarchical Models for Panel Data

As we have discussed, reasonably precise inferences about the MNP model parameters, particularly correlation and covariance parameters, may require a large number of observations. It is rare to observe only one economic agent making choices between the same set of alternatives on many different occasions. Large samples are obtained in MNP applied work by observing the choices made by a large number of agents. Frequently, these choices are observed in a panel setting in which a large number of entities are observed for a relatively short period of time. In these panel settings, it is important that the MNP model be able to accommodate differences between entities. From this point on, we will refer to the economic entities as households even though they could just as easily be firms.

Differences between households in model coefficients or "heterogeneity" has received a good deal of attention in the econometric literature.. The standard approach to this problem of accommodating heterogeneity with large N and small T is to use a random coefficient model. In the MNP literature, we typically see heterogeneity modeled as a random coefficient model for the intercepts (c.f. Borsch-Supan and Hajivassilliou [1992]). There is no particular reason to believe that heterogeneity is restricted to the intercepts (for example, it is entirely reasonable that different households might display different sensitivities to choice characteristics such as price). In addition, a distinction is often made between "observable" heterogeneity (differences which are linked to observable attributes of households) and "unobservable" heterogeneity which is only revealed by choice behavior. Therefore, it is imperative that our approach to modeling heterogeneity accommodate differences in all MNP regression coefficient parameters as well as incorporate observable and unobservable heterogeneity.

7.1 A Hierarchical Approach to Modeling Heterogeneity

To fix the notation, we consider a panel of N households observed over T_h periods for each household.

$$w_{h,t} = X_{h,t}\beta_h + \varepsilon_{h,t} \qquad \varepsilon_{h,t} \sim \text{iid}\, N(0,\Sigma)$$

$$h = 1,\ldots,H \qquad t = 1,\ldots,T_h$$

We model the household heterogeneity via an additional regression model.

$$\beta_h = \Delta z_h + v_h \qquad v_h \sim N\left(0, V_\beta\right)$$

z_h contains a vector of d household characteristic ("demographic") variables. Thus, each MNP regression coefficient is related to a vector of characteristics and an unobservable component v. The unobservable component has a general covariance structure over households. Since we typically do not want to fix Δ and V_β as known, we introduce priors for these common parameters. This is a good example of a hierarchical Bayesian model in which the sampling model and set of priors is built up from a series of conditional distributions.

The complete model is specified as follows:

i. $y_{h,t} | w_{h,t}$

ii. $w_{h,t} | X_{h,t}, \beta_h, \Sigma$

iii. $\beta_h | z_h, \Delta, V_\beta$

iv. $\Sigma | \nu, V$

v. $\Delta | \bar{\Delta}, A, V_\beta$

vi. $V_\beta | \nu_\beta, V_0$

$\bar{\Delta}, A, \upsilon_\beta, V_0$ are hyper-parameters of the priors for Δ and V_β. The conditionals i and ii specify what most consider the "sampling" model. Conditional iii. specifies what econometricians call the "random coefficient" model. In our Bayesian hierarchical approach, the conditionals iii-vi specify a joint prior over the set of $\{\beta_h\}$. Rossi, McCulloch and Allenby (1995) discuss the specific form of these priors using the conditionally conjugate families.

The prior on V_β is especially important in determining how much information is shared or "borrowed" across households in performing inference about the $\{\beta_h\}$. If the prior for V_β is tight around a small value, then there will be extensive shrinkage and the $\{\beta_h\}$ will differ little. More diffuse priors on V_β will induce less shrinkage. It is important to note that this prior must be proper in order for the joint posterior in this model to be proper. Thus, even with very diffiuse settings, the prior on V_β must be influential. By making this a proper prior, we are asserting that there is some commonality among the $\{\beta_h\}$.

Given the basic framework outlined for the fixed coefficient MNP model outlined in sections 3 above, it is a simple matter to elaborate the Gibbs sampler to include conditional posterior draws of the Δ and V_β which are in well-known multivariate normal and Wishart form (see Rossi, et al [1995] for details). Our experience with this more elaborate sampler is that the sampler converges rapidly to a stationary distribution but that the draw sequences can be highly autocorrelated, necessitating long runs for accurate results.

In some problems, it may be useful to make inferences about the draw of β_h for a specific household. In the Bayesian hierarchical approach, we require the marginal posterior distribution of β_h.

$$p(\beta_h|\text{data}) \propto \int p(\beta_h, \Delta, V_\beta|\text{data})d\Delta dV_\beta$$

Fortunately, we can easily marginalize on β_h using the sequence of Gibbs draws produced as a by-product of our procedure. Rossi et al (1995) make explicit use of these household-level parameter inferences to solve various targeted marketing problems which involve customizing promotional offers at the household level. The β_h draws can also be used as the basis of an informal diagnostic procedure for the form of the prior or mixing distribution. The assumption $\beta_h \sim N(\Delta z_h, V_\beta)$ is only the form of the prior for each household. The individual household data can shape the posterior for household h to a different form. If we lump together all draws across all households and look at this distribution, we can assess whether the normal prior is appropriate to characterize the distribution of β_h over households. For example, if we see a multi-modal distribution, this might suggest that we should investigate more flexible priors.

In a classical random coefficient model, the likelihood is averaged over the mixing distribution and inference is only made about the common parameters of the mixing distribution. On the other hand, the Bayesian approach combines the smoothing advantages of the frequentist random effects model with the richness of the fixed effects approach. In the Bayesian approach, there is no real distinction between fixed and random effects only between independence priors ("fixed" effects) and dependent priors (hierarchical or random coefficient models).

7.2 Extensions

7.2.1 Normal Mixtures

The normal mixing model used in section 7.1 as the first stage of the hierarchical prior can be criticized as insufficiently flexible. In many situations, we might want to specify a prior structure that would allow for some grouping or clustering of households into more homogenous sub-populations. One way of achieving this would be to specify a mixture of normals as the prior. To illustrate how this might be achieved, we will simplify the model of section 7.1 to remove the household characteristics vector, z_h. In this model, we would simply have a normal prior with a fixed mean vector, $\beta_h \sim N\left(\bar{\beta}, V_\beta\right)$. We can replace this normal distribution with a mixture of normals:

$$p\left(\beta_h \middle| \lambda, \left\{\bar{\beta}_j\right\}, \left\{V_{\beta,j}\right\}\right) = \sum_{j=1}^{J} \lambda_j \varphi\left(\beta_h \middle| \bar{\beta}_j, V_{\beta,j}\right) \qquad \sum_{j=1}^{J} \lambda_j = 1$$

We would have to introduce a prior for the λ_j mixture probabilities, and it would be most convenient to use a standard natural conjugate Dirichlet prior. The mixture of normals model could be easily handled in the Gibbs sampler via the introduction of latent indicator variables which switch on for each of the J components in the mixture. The indicator variables would have a conditional multinomial distribution. Given the component indicator variables, inference about the parameters of each component in the mixture can be done using standard normal model theory.

7.2.2 Multiperiod Probit Models

In many panel applications, the assumption of independence of the model errors across time for the same households is questionable. It is straightforward to extend the MNP samplers to handle error terms which follow an AR(p) process. To illustrate this idea, consider a binomial multiperiod probit model. Geweke, Keane and Runkle (1994) consider a Gibbs sampling approach to the multinomial multiperiod probit model and conduct extensive comparisons with the methods of simulated moments and method of simulated likelihood.

The latent variables set-up for the binomial multiperiod probit is given as follows:

$$w_{h,t} = x_{h,t}{}' \beta_h + \varepsilon_{h,t} \qquad \text{where } \varepsilon_{h,t} = \Gamma_p(B)\varepsilon_{h,t} + u_{h,t}; \;\; \text{Var}\left(u_{h,t}\right) = 1$$

The AR polynomial is parameterized by p autoregressive coefficients which we denote by the vector ϕ. The $u_{h,t}$ are assumed to be independent across time and households. We must introduce priors for the $\varepsilon_{h,0}$ and ϕ. To implement the Gibbs sampler, we must modify our strategy for drawing w and append two sets of conditionals to the Gibbs structure. This Gibbs sampler consists of the following set of conditionals:

$$w|\beta, \varphi, \varepsilon_0, y, X$$

$$\beta|w, \varphi, \varepsilon_0, X$$

$$\varphi|w, \beta, \varepsilon_0, X$$

$$\varepsilon_0|w, \varphi, \beta, X$$

Here β, ε_0 ,w,y and X are stacked vectors (matrices) of household parameters and data. The draw of w proceeds using the same strategy as before except that the autocorrelation structure changes the univariate normal conditional distribution of $w_{h,t}|w_{h,t-1}$. In addition, the conditional posterior of β is computed as before except on the orthogonalized regression system (pre-multiplied by the Cholesky root of the correlation structure which is available since we are conditioning on ϕ).

8. Conclusions

In this chapter, we illustrate how Bayesian inference can be achieved for a number of variants of the multinomial probit model. In particular, we consider various informative and non-informative priors on the model parameters, accommodating heterogeneity of various forms, non-normal mixing distributions, and multiperiod models. All of these situations can be handled via a Gibbs sampling strategy in which a Markov chain is constructed with the posterior distribution as its invariant distribution. Extensions of the basic model are handled in a unified framework by appending additional distributions to the base Gibbs sampler.

Our experience to date with the MNP model with both actual and simulated data suggests that there is much promise in this line of research. All of the data sets we have analyzed strongly support a rejection of the IIA property which is at the core of the multinomial logit model. The MNP model provides a great deal of flexibility at the expense of a complicated and high-dimensional parameterization. Our own experience is that it is difficult to make precise inferences about the covariance structure of the latent variable errors. This suggests that future successful modeling approaches will rely on restricted MNP models or informative priors.

Finally, Bayesian methods offer an alternative to the recent classical methods of simulated moments, simulated scores and simulated maximum likelihood. These methods provide the basis for finite sample inferences at approximately the same order of computation demands as these classical alternatives. In addition, our methods provide a natural and flexible method for modeling heterogeneity and provide household level parameter inferences should these be needed. The chief practical concern in the application of our methods is the rate of convergence of the Gibbs sampler. Our experience is that convergence is rapid enough for reliable practical application.

References

Barnard, J. , R. E. McCulloch, and X. Meng (1995), "Shrinkage Priors for Covariance Matrices," working paper, University of Chicago.

Borsch-Supan, A. and V. Hajivassilious (1990), "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*, Cambridge, NBER, 79-107.

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

Dansie, B. (1985), "Parameter Estimability in the Multinomial Probit Model," *Transportation Research -B*, 19B:6, 526-28.

Gelman, A. and D. Rubin (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*7, 457-511.

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., M. Keane, and D. Runkle (1994), "Alternative Computational Approaches to Statistical Inference in the Multinomial Probit Model", *Review of Economics and Statistics* , 76, 609-632.

Geweke, J., M. Keane, and D. Runkle (1994), "Statistical Inference in the Multinomial, Multiperiod Probit Model," forthcoming, *Journal of Econometrics.*

Hajivassiliou, V. and P. Ruud (1994), "Classical Estimation Methods for LDV Models Using Simulation," in Engle and McFadden, eds, *Handbook of Econometrics*, Amsterdam: North Holland, 2384-2438.

Keane, M. (1992), "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics* 10, 193-200.

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

McCulloch, R. E. , N. G. Polson, and P. E. Rossi (1993), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," working paper, Graduate School of Business, University of Chicago.

Nobile, A. (1995), "A Hybrid Markov Chain for Bayesian Analysis of the Multinomial Probit Model," working paper, Institute of Statistics and Decision Sciences, Duke University.

Ritter, C. and M. A. Tanner (1992), "Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy-Gibbs sampler," *Journal of the American Statistical Association*, 48, 276-279.

Rossi, P., R. McCulloch, and G. Allenby (1995), "The Value of Purchase History Data in Target Marketing," working paper, University of Chicago.

Tanner, M and W. Wong (1987), "The Calculations of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82, 528-549.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, School of Statistics, University of Minnesota.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley and Sons.

Zellner, A. and P. E. Rossi (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics* 25, 365-393.