Maximum likelihood for the multinomial probit model

Ranjini NATARAJAN*[a], Charles E. McCULLOCH[b] and Nicholas M. KIEFER[c],

[a] *School of Operations Research and Industrial Engineering*

[b] *Biometrics Unit and Statistics Center*

[c] *Department of Economics*

Cornell University, Ithaca, NY 14853-3801, USA.

Abstract

In this paper we develop a framework to perform maximum likelihood estimation of the multinomial probit model using a Monte Carlo EM algorithm. Our method includes a Gibbs step. This approach is different from standard likelihood procedures in that it does not involve direct evaluation and maximization of the observed data likelihood. Instead, we take advantage of the underlying continuum to simplify calculations. Estimation of the asymptotic standard errors of these ML estimates is described. We also develop extensions of this Monte Carlo EM method for analyzing multi-period and non-normal data. The computations are illustrated through three examples.

Key Words: Gibbs Sampling, Maximum likelihood estimation, Monte Carlo EM, Observed Information, Panel Data.

Classification Code: **C**13, **C**15.

*Corresponding Author: Ranjini Natarajan,
288 ETC, School of Operations Research and Industrial Engineering,
Cornell University, Ithaca, New York 14853-3801, USA.
Telephone: (607) 255-9139
Fax: (607) 255-9129
E-mail: ranjini@orie.cornell.edu

BU-1286-M

May 1995

1 Introduction

The multinomial probit (MNP) model for unordered categorical data has its roots in the biometrics and econometrics literature (Ashford and Sowden 1970, Hausman and Wise 1978 and Daganzo 1980). The appeal of this model is that arbitrarily complicated covariance structures (including multi-period data) can be modeled quite naturally, unlike the multinomial logit model. However, despite its broad applications in discrete economic choice behavior, usage has been limited by the computational burden associated with estimating the parameters. Evaluation of the likelihood function for this model requires a method to compute multi-normal orthant probabilities. This calculation is difficult unless the dimension of the multi-normal distribution is less than five, or the error variance matrix has a special structure (Hajivassiliou, McFadden, Ruud, 1992).

Previous work on a likelihood analysis of the MNP model focuses on numerically maximizing a simulation-based estimate of the likelihood function (Lerman and Manski 1981, Geweke 1989, Stern 1992, Börsch-Supan and Hajivassiliou 1993). This simulated maximum likelihood approach suffers from two severe drawbacks. First, it deals with likelihood calculations which are typically computationally difficult compared to log-likelihood calculations. Often this drawback manifests itself in the form of numerical instabilities, sensitivity to starting values and poor convergence properties, especially for larger problems. Second, there appears to be a misplaced focus on the accuracy and precision of the simulator, rather than efficient maximization using such a simulator. For problems with a large number of parameters, efficient maximization should be as important a concern.

Some work has been done on developing alternative estimation methods (McFadden 1989, Hajivassiliou and McFadden 1990, Keane 1993, McCulloch and Rossi 1994). McFadden (1989) proposed the method of simulated moments (MSM) which involves substitution of simulated orthant probabilities into moment conditions. The computational effort required for this method grows quite rapidly relative to the size of the problem, especially for multi-period models. Keane (1993) suggested a computationally feasible variant of this MSM approach which factors the choice probabilities into transition probabilities. McCulloch and Rossi (1994) describe a Bayesian analysis of the MNP model using very diffuse, but proper priors. Their justification for using proper priors rests on the fact that improper priors do not necessarily lead to proper posterior distributions for this class of models (see Natarajan and McCulloch (1995) for a proof for a special case).

This paper describes maximum likelihood estimation in the MNP model via a Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990). Our motivation for using EM is three-fold: (i) it circumvents direct evaluation of the likelihood function which is a tremendous computational advantage; (ii) the iterates are automatically confined to lie in the parameter space; (iii) in practice, it has been found to converge from a wider range of starting values than other iterative maximization routines. A Monte Carlo implementation of Louis' (1982) method is developed to estimate the asymptotic variance matrix of the MLE. We compare our MCEM approach with the smooth simulated maximum likelihood (SSML) approach advocated by Börsch-Supan and Hajivassiliou (1993), which they show to be superior to extant methods. We demonstrate through a simulated example that MCEM converges to the MLE more quickly and accurately than SSML. The computational time required for each of these methods to converge is also reported for some larger problems.

Extensions of this MCEM method for analyzing multi-period models and non-normal data are developed. We show that our approach can be adapted very naturally to accommodate complicated panel models with any number of fixed effects and arbitrary variance structures. While the SSML approach can also be generalized to handle panel data, it does so by directly estimating the high-dimensional integrals that result from repeated observations on each individual. This method can become computationally intensive and inefficient for large problems with several discrete choices and time periods. Lastly, an important advantage of the MCEM framework is the fact that it is not dependent on the normality assumption of the MNP model. In fact it can be readily adapted for other limited dependent variable models (multinomial logit, Tobit regression). This has not been previously possible. Current likelihood procedures for the MNP model are intrinsically dependent on distributional assumptions.

This paper is organized as follows. The MNP model (for a single time period) is formulated in Section 2. By way of notation, we will use boldface characters to denote vectors and $'$ for the transpose operator. In Section 3 we describe maximum likelihood estimation for this model. Estimation of the asymptotic standard errors is discussed in Section 4. In Section 5 the computations are illustrated through two examples. The first example is a simulated 3 choice data set with a single fixed effect. We show that MCEM is more accurate and efficient than SSML. The second example is actual customer data on the quantity demanded of several menu items at a medium-priced family restaurant (Kiefer, Kelly and Burdett, 1994). In Section 6 we demonstrate that MCEM can be easily modified to handle multi-period data as well as non-normal models. We use this multi-period MCEM procedure to study consumer panel data on the household purchase of peanut butter (ERIM panels, A. C. Nielsen).

2 The Model

Assume that N agents/individuals choose among a set of c choices. The observed data is a multinomial vector $\mathbf{w}_i = \{w_{i1}, \ldots, w_{ic}\}'$ for every agent i. Each component of $\mathbf{w}_i$ is binary and specifically:

$$w_{ij} = I(\text{agent } i \text{ chooses choice } j),$$

where $I(.)$ is the indicator function. Further, the set of c choices are discrete so that $\sum_{j=1}^{c} w_{ij} = 1$. The MNP model arises by postulating the existence of a latent continuum $\mathbf{u}_i = \{u_{i1}, \ldots, u_{ic}\}'$ which generates the observed $\mathbf{w}_i$ in the following manner:

$$w_{ij} = I\left(u_{ij} = \max_{k} \{u_{ik}\}\right), \tag{1}$$

and

$$\mathbf{u}_i \sim \mathcal{N}_c(X_i^* \beta^*, \Omega^*), \tag{2}$$

where X_i^* is a known $c \times p$ design matrix of exogenous variables, β^* the unknown $p \times 1$ vector of fixed effects, and Ω^* a $c \times c$ variance matrix.

It is typical to regard the $\mathbf{u}_i$ as the un-measured utility or value of the c choices to the i^{th} individual. Thus, model (1) suggests that an agent picks the choice that has the largest utility to them. It is unimportant whether we actually believe in these underlying $\mathbf{u}_i$ or simply use it as a mechanism for estimation. However, in practice there are numerous applications where the existence of the underlying continuum can be easily justified (marketing, economics, biometrics, public health).

The choice model as stated in (1) and (2) is not identified (Dansie, 1985). In order to achieve identification it is conventional to re-formulate it in terms of the relative differences of the utilities from some baseline utility. Define $y_{ij} = u_{ij} - u_{ic}, \forall j$. It is easy to see that $\mathbf{y}_i = \{y_{i1}, \ldots, y_{ic-1}\}'$ satisfies:

$$\mathbf{y}_i \sim \mathcal{N}_{c-1}(X_i \beta, \Omega) \tag{3}$$

where X_i, β, Ω are the appropriate transformations of X_i^*, β^*, Ω^*. Further, (1) can be re-expressed as

$$w_{ij} \;=\; I\left(y_{ij} \;=\; \max_k\{y_{ik}\}\right). \tag{4}$$

Since the scale of the relative utilities y_{ij} is indeterminate, we set the first diagonal element of Ω equal to 1 (Geweke, Keane and Runkle, 1994).

2.1 The Estimation Problem

The focus of this paper is on maximum likelihood estimation of β and Ω. In practice there can be several other quantities of interest, for example, the choice probabilities, gradients of these probabilities, etc. The MNP likelihood is given by:

$$L(\beta,\,\Omega\,;\mathbf{w}_1,\,\dots,\mathbf{w}_N) \;=\; \prod_{i=1}^{N}\prod_{j=1}^{c}\pi_{ij}^{w_{ij}} \tag{5}$$

where π_{ij} is the probability that individual i chooses choice j. For each i, these π_{ij} satisfy the constraints:

$$\pi_{ij} \;>\; 0,\; \sum_{j=1}^{c}\pi_{ij} \;=\; 1,$$

and are given by:

$$\begin{aligned}\pi_{ij} \;&=\; P\left(y_{ij} \;=\; \max_k\{y_{ik}\}\right),\\ &=\; \int \phi_{c-1}\left(X_i\,\beta,\,\Omega\right)\, I\left(y_{ij} = \max_k\{y_{ik}\}\right) d\mathbf{y}_i,\end{aligned}$$

where $\phi_{c-1}(.)$ is the $(c\,-\,1)$ dimensional multi-normal density function. The above integral does not have a closed form except for special cases of Ω.

Thus, the estimation problem is complicated since the likelihood is typically not available in closed form. Further, if the number of discrete choices is greater than five, numerical integration estimates of the π_{ij} are not a computationally feasible or accurate option (Hajivassiliou, McFadden and Ruud, 1992). Current techniques to estimate the MLE focus on obtaining a simulation-based estimate $\hat{\pi}_{ij}$ of the choice probabilities π_{ij}. Numerical routines are then used to maximize the estimate of the likelihood function formed by inserting $\hat{\pi}_{ij}$ into (5). Several estimators of π_{ij} have been proposed (Lerman and Manski 1981, Geweke 1989, Stern 1992, Börsch-Supan and Hajivassiliou

1993). However, despite the extensive results published on the accuracy and precision of these simulators, there have been very few results on the accuracy of the resulting approximations to the MLE. The examples described in Section 5.1 suggest that if the goal is maximum likelihood estimation, it is not adequate to only focus on good estimates of the likelihood, but in fact efficient maximization should be as important a concern.

We develop a Monte Carlo EM approach to estimate the MLE of β and Ω. This approach circumvents direct evaluation and maximization of the observed likelihood by taking advantage of the latent structure of the model. We demonstrate through three examples that our approach is computationally feasible for any number of choices, fixed effects parameters and arbitrarily complicated covariance structures.

3 Maximum Likelihood Estimation

In this section we describe maximum likelihood estimation using a Monte Carlo EM algorithm. The Expectation Maximization algorithm (Dempster, Laird and Rubin, 1977) is a powerful numerical tool used for computing ML estimates in standard incomplete data problems. The basic premise is that the maximization based on the observed (incomplete) data is computationally intractable. However, by augmenting the observed data, the hard maximization can be reduced into a sequence of easier problems. We first review the EM algorithm in general and then describe the specifics for the MNP model.

3.1 Review of the EM algorithm

In the usual EM terminology let $\mathbf{z} = \{\mathbf{y}, \mathbf{w}\}$ denote the complete data, where $\mathbf{w}$ is observed and $\mathbf{y}$ is missing data. We assume that $\mathbf{z}$ is indexed by a d-dimensional parameter θ, and the goal is to find the MLE of θ. If $\mathbf{z}$ were observed, the objective would be to maximize:

$$\mathcal{L}(\theta\,;\,\mathbf{z}) = \ln\left[\mathbf{z} \mid \theta\right]$$

where $[\,.\,]$ denotes probability density or mass functions. However, since only the $\mathbf{w}$ are observed, we need to maximize:

$$\mathcal{L}(\theta\,;\,\mathbf{w}) = \ln\left[\mathbf{w} \mid \theta\right]$$

$$= \ln \int [\mathbf{y},\, \mathbf{w} \mid \theta]\, d\mathbf{y}.$$

It is the integration which can make the maximization of the observed data log-likelihood tedious, even when maximizing the complete data log-likelihood is trivial. The EM algorithm maximizes $\mathcal{L}(\theta\,;\,\mathbf{w})$ by iteratively maximizing $E\,[\mathcal{L}(\theta\,;\,\mathbf{z}) \mid \mathbf{w}]$. Each iteration has 2 steps: an Expectation-step and a Maximization-step. The $(m+1)^{st}$ E-step computes:

$$Q(\theta \mid \theta^{(m)}) \;=\; E\,[\mathcal{L}(\theta\,;\,\mathbf{z})]$$

where the expectation is with respect to the conditional density of the missing data given the observed data, namely, $[\mathbf{y} \mid \mathbf{w},\, \theta^{(m)}]$. The $(m+1)^{st}$ M-step then finds $\theta^{(m+1)}$ to maximize $Q(\theta|\theta^{(m)})$. Although this algorithm works quite generally for any model, it is particularly useful when the complete data are from an exponential family, since the E-step merely reduces to finding the complete-data sufficient statistics.

Sometimes the computations required for the E-step are hefty. In such cases a Monte Carlo estimate can be obtained by estimating $Q(\theta \mid \theta^{(m)})$ by

$$\frac{\sum_{r=1}^{R} \mathcal{L}(\theta\,;\,\mathbf{y}^{(r)},\, \mathbf{w})}{R},$$

where $\mathbf{y}^{(r)} \sim [\mathbf{y} \mid \mathbf{w},\, \theta^{(m)}]$, $r = 1, ..., R$. This leads to a Monte Carlo EM method (Wei and Tanner 1990).

3.2 MCEM for the MNP Model

We use the EM algorithm with the following definitions. We regard the missing data as the vector of relative utilities $\{\mathbf{y}_i,\, i = 1, ..., N\}$. Once the $\mathbf{y}_i$ are assumed known, the observed $\mathbf{w}_i$ are degenerate. Hence, the complete data is simply the multi-normal vector $\{\mathbf{y}_i,\, i = 1, ..., N\}$. It is this fact that allows EM to circumvent direct evaluation of the likelihood function. We will see later that the observed $\mathbf{w}_i$ only comes into play in terms of defining the appropriate region of truncation for $\mathbf{y}_i$. The complete data log-likelihood is given by:

$$\begin{aligned} \mathcal{L}(\beta, \Omega\,;\, \mathbf{y}_1, ..., \mathbf{y}_N) &\propto -\frac{N}{2} \ln|\Omega| - \frac{1}{2} \sum_{i=1}^{N} \mathbf{e}_i'\, \Omega^{-1}\, \mathbf{e}_i \\ &= -\frac{N}{2} \ln|\Omega| - \frac{1}{2}\, tr\left(\Omega^{-1} \sum_{i=1}^{N} \mathbf{e}_i\, \mathbf{e}_i'\right) \end{aligned}$$

where $\mathbf{e}_i = \mathbf{y}_i - X_i\,\beta$. The E-step is conceptually simple. For each i, we need to calculate:

$$\begin{aligned} Q_i(\beta) &= E\left[\mathbf{e}_i\,\mathbf{e}_i' \mid \mathbf{w}_i\right], \\ &= Var\left[\mathbf{y}_i \mid \mathbf{w}_i\right] + E\left[(\mathbf{y}_i - X_i\,\beta) \mid \mathbf{w}_i\right]\, E\left[(\mathbf{y}_i - X_i\,\beta) \mid \mathbf{w}_i\right]'. \end{aligned}$$

The M-step is also straightforward. We maximize

$$-\frac{N}{2}\ln|\Omega| - \frac{1}{2}\,tr\left(\Omega^{-1}\sum_{i=1}^{N} Q_i(\beta)\right)$$

with respect to β and the $\left(\frac{c\,(c-1)}{2} - 1\right)$ parameters in Ω. It is well known that the joint maximizing values of β and Ω are not in closed form. Thus, we adopt Meng and Rubin's (1993) suggestion of replacing this M-step with 2 conditional M-steps; the first is a maximization with respect to β conditional on the elements of Ω, and the second is a maximization over the unknown elements of Ω conditional on the updated value of β. This leads to a Monte Carlo Expected Conditional Maximization (MCECM) algorithm. However, in the rest of this paper, we will continue to refer to our approach as a Monte Carlo EM (MCEM) method for brevity. The conditional ML estimate of β is simply the generalized least squares estimator:

$$\beta^{(m+1)} = \left[\sum_{i=1}^{N} X_i'\,\Omega^{(m)-1}\,X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'\,\Omega^{(m)-1}\,E\left[\mathbf{y}_i \mid \mathbf{w}_i\right]\right],$$

while the conditional ML estimate of Ω is obtained by maximizing

$$-\frac{N}{2}\ln|\Omega| - \frac{1}{2}\,tr\left(\Omega^{-1}\sum_{i=1}^{N} Q_i(\beta^{(m+1)})\right) \tag{6}$$

with respect to the parameters in Ω. Although the maximization of (6) requires an iterative procedure, it is a standard calculation and would need to be performed even if the $\mathbf{y}_i$ were observed. Non-linear functions such as (6) are relatively well-studied and there are several extremely efficient algorithms to perform the maximization (Jennrich and Schluchter 1986).

Thus, we have shown that the only additional computations required for ML estimation of discrete choice data is the calculation of $E\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$ and $Var\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$. We are now ready to make a formal statement of the EM algorithm. The superscripts in parentheses on $Var\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$ and $E\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$ indicate that current parameter values have been substituted.

3.3 The Algorithm

Step 0) Obtain starting values $\beta^{(0)}$ and $\Omega^{(0)}$. Set counter $m = 0$.

Step 1) (E-step) For each i, calculate:

$$E^{(m)}[\mathbf{y}_i \mid \mathbf{w}_i], \text{ and } Var^{(m)}[\mathbf{y}_i \mid \mathbf{w}_i].$$

Step 2a) (Conditional M-step 1) Set:

$$\beta^{(m+1)} = \left[\sum_{i=1}^{N} X_i' \Omega^{(m)-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' \Omega^{(m)-1} E^{(m)}[\mathbf{y}_i \mid \mathbf{w}_i]\right]$$

Step 2b) (Conditional M-step 2) Maximize:

$$-\frac{N}{2} \ln|\Omega| - \frac{1}{2} tr\left(\Omega^{-1} \sum_{i=1}^{N} Q_i(\beta^{(m+1)})\right)$$

over the unknown elements of Ω to obtain $\Omega^{(m+1)}$.

Step 4) If convergence is reached set $\hat{\beta}_{MCEM} = \beta^{(m+1)}$ and $\hat{\Omega}_{MCEM} = \Omega^{(m+1)}$; else increment counter m by one and return to Step 1).

Most of the computational effort is expended in computing the conditional means and variances of $\mathbf{y}_i$ given the observed data $\mathbf{w}_i$. For small problems, and simple covariance structures Ω these can be computed using direct numerical integration. However, for more complicated models we propose the use of the Gibbs sampler (Geman and Geman 1984) to estimate them. More details on this will be discussed in Section 3.5. An important consideration in implementing Monte Carlo EM is the monitoring of convergence. We will now discuss this issue.

3.4 Convergence of MCEM

The convergence of MCEM can be monitored by plotting the parameter value at each iteration versus the iteration number. After a certain number of iterations the plot will reveal random fluctuation about the maximum likelihood estimator, due to the randomness introduced by the Monte Carlo E step (see Figures 1 and 2). At this point one may either terminate the algorithm, or continue with a large number of Gibbs samples to decrease the Monte Carlo variability. Chan and Ledolter (1995) provide a stopping criterion as well as rules for selecting the appropriate Monte Carlo sample size.

From our experience on both simulated and real data sets we have found the convergence of MCEM to be fast. For the examples we studied, we also found MCEM to be extremely robust to the choice of starting value. We will now discuss the estimation of the expectations involved in the E-step.

3.5 The Gibbs Sampler

The computational difficulties associated with calculating the moments of the conditional distribution $[\mathbf{y}_i \mid \mathbf{w}_i]$ can be burdensome, since it involves integrals similar to those that appear in the likelihood function. We propose the use of the Gibbs sampler to provide a simulation-based estimate of these moments. This application of the Gibbs sampler is rather unusual in that it is used to solve for ML estimates directly rather than within a Bayesian framework. McCulloch (1994) developed a similar Monte Carlo EM algorithm for the simple two choice problem with random effects.

The efficient implementation of the Gibbs sampler rests on the fact that fast acceptance-rejection algorithms exist to generate from truncated univariate normal distributions (e.g. inverse transform method, Devroye 1986). In order to generate a sample from the density $[\mathbf{y}_i \mid \mathbf{w}_i]$ using the Gibbs sampler, we need to cycle through the full conditional specifications

$$[y_{ij} \mid \mathbf{y}_{i(-j)}, \mathbf{w}_i]$$

where $\mathbf{y}_{i(-j)} = \{y_{i1}, \dots, y_{ij-1}, y_{ij+1}, \dots, y_{ic}\}'$. Using standard results on normal theory, it can be shown that these conditional densities are simply univariate truncated normal distributions (Searle, Casella and McCulloch 1992). Here is an outline of how the Gibbs sampler is used to generate a sample of $\mathbf{y}_i$ from the conditional distribution of $[\mathbf{y}_i \mid \mathbf{w}_i]$ for a fixed i.

a) For each $j = \{1, \dots, (c-1)\}$ calculate:

$$\begin{aligned} \sigma^2_{j\,|\,-j} &= Var\left[y_{ij} \mid \mathbf{y}_{i(-j)}\right], \\ \beta_{j\,|\,-j} &= Cov\left[y_{ij}, \mathbf{y}_{i(-j)}\right]. \end{aligned}$$

These are standard calculations for normally distributed variates.

b) For each $j = \{1, \dots, (c-1)\}$ calculate:

$$\mu_{ij\,|\,i(-j)} = E\left[y_{ij} \mid \mathbf{y}_{i(-j)}\right],$$

$$= x_{ij}\beta + \beta'_{j\,|\,-j}(\mathbf{y}_{i(-j)} - X_{i(j)}\beta),$$

where $X_{i(j)} = X_i$ with row j deleted and x_{ij} is the j^{th} row of X_i.

c) Simulate y_{ij} from a truncated normal distribution with mean $\mu_{ij\,|\,i(-j)}$ and standard deviation $\sigma_{j\,|\,-j}$. If $w_{ij} = 1$, simulate y_{ij} truncated above $\max\{\mathbf{y}_{i(-j)}\}$; else simulate y_{ij} truncated below $\max\{\mathbf{y}_{i(-j)}\}$.

Repeat Steps b) and c) a large number of times, say M, to obtain $\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, ..., \mathbf{y}_i^{(M)}$. Discard a suitable number from the beginning of the sequence, *nburn*, and then accept every *nskip*th one to form a sample of size *nrep*. This sample is then used to estimate $E\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$ and $Var\left[\mathbf{y}_i \mid \mathbf{w}_i\right]$. There are no hard and fast rules on the choice of *nburn*, *nrep* and *nskip*. On account of the iterative nature of EM and the desire to take as few Gibbs samples as possible at the beginning of EM, we decided to let these numbers depend on the iteration of EM; i.e., later iterations do more Gibbs sampling.

It is evident that this approach can accommodate arbitrarily complicated covariance structures since they only affect Step a), which is performed only once before initiating the Gibbs chain. We will now discuss the calculation of standard errors for the ML estimates of β and Ω.

4 Standard Error Calculations

In this section we describe a Monte Carlo approach to Louis' method (1982) to estimate the asymptotic standard errors of the MLE of the MNP model. Guo and Thompson (1992) outline a similar method for genetic models.

4.1 Louis' Method

Louis developed a technique to compute the observed information matrix within the EM framework. It requires computation of the complete data gradient vector and second derivative matrix and can be embedded quite simply in the EM iterations. In order to describe his method we use the notation developed in the review of EM in Section 3.1. Louis proved that the observed information matrix $I_w(\theta)$ satisfies the following identity:

$$I_w(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ln[\mathbf{z}\mid\theta]\,\middle|\,\mathbf{w}\right] - Var\left[\frac{\partial}{\partial\theta}\ln[\mathbf{z}\mid\theta]\,\middle|\,\mathbf{w}\right],$$

$$= \quad -E\,[H(\mathbf{z};\theta)\,|\,\mathbf{w}] - Var\,[S(\mathbf{z};\theta)\,|\,\mathbf{w}]\,, \tag{7}$$

where S and H are the complete data score vector and Hessian matrix respectively. Of course, they need be evaluated only on the last iteration of EM at the MLE.

Despite the conceptual simplicity and elegance of Louis' identity, this method has not been used extensively due to computational difficulties in evaluating the expectations involved. A computationally feasible variant of Louis' identity can be obtained by replacing all the expectations in (7) by their Monte Carlo estimates, in the following manner:

1) Generate $\mathbf{y}^{(r)} \sim [\mathbf{y}\,|\,\mathbf{w},\,\theta],\;\; r = 1, \ldots, R.$

2) Replace each term in (7) by its Monte Carlo estimate, e.g., replace the first term by

$$-\frac{1}{R}\sum_{r=1}^{R} H(\mathbf{y}^{(r)},\,\mathbf{w};\theta).$$

We use this Monte Carlo implementation to estimate the observed information matrix for the MNP model. Expressions for the elements of S and H are given in Appendix 2. Although the matrix manipulations look messy, they are all one-time calculations and we can exploit the general form of Ω to obtain simple expressions (Jennrich and Schluchter, pg 813, 1986). We will now demonstrate the calculations involved through two examples.

5 Examples

In this section we provide two examples to illustrate the feasibility of MCEM. The first example is a simple 3 choice simulated data set. The second is actual customer data on the quantity demanded of several menu items.

5.1 Variance Components Setup

The first example is intended to serve as a comparison with the SSML approach of Supan and Hajivassiliou (1993). Consider a simple 3-choice model with a single fixed effect and a low order factor structure for the error variance matrix. More formally we consider the following latent structure for the relative utilities:

$$y_{ij} \quad = \quad x_{ij}\beta + u_i \, + \, \epsilon_{ij},\; j \, = \, 1,\, 2,$$

$$\begin{aligned} u_i &\sim \mathcal{N}(0,\, \theta), \\ \epsilon_{ij} &\sim \mathcal{N}(0,\, 1). \end{aligned} \tag{8}$$

Hence Ω is characterized by one parameter θ and is given by:

$$\Omega = \begin{pmatrix} 1+\theta & \theta \\ \theta & 1+\theta \end{pmatrix}.$$

The contribution to the observed data likelihood by the i^{th} individual is $\prod_{j=1}^{3} \pi_{ij}^{w_{ij}}$, where π_{ij} are the choice probabilities and are given by:

$$\pi_{i1} = \int_{-x_{i1}\beta}^{\infty} \frac{\phi\left(\frac{u_1}{\sqrt{1+\theta}}\right)}{\sqrt{1+\theta}} \Phi\left(\frac{(x_{i1}-x_{i2})\beta + \frac{u_1}{(1+\theta)}}{\sqrt{\frac{1+2\theta}{1+\theta}}}\right) du_1,$$

$$\pi_{i2} = \int_{-x_{i2}\beta}^{\infty} \frac{\phi\left(\frac{u_1}{\sqrt{1+\theta}}\right)}{\sqrt{1+\theta}} \Phi\left(\frac{(x_{i2}-x_{i1})\beta + \frac{u_1}{(1+\theta)}}{\sqrt{\frac{1+2\theta}{1+\theta}}}\right) du_1,$$

and

$$\pi_{i3} = \int_{-\infty}^{-x_{i1}\beta} \frac{\phi\left(\frac{u_1}{\sqrt{1+\theta}}\right)}{\sqrt{1+\theta}} \Phi\left(\frac{-x_{i2}\beta - \frac{\theta u_1}{(1+\theta)}}{\sqrt{\frac{1+2\theta}{1+\theta}}}\right) du_1,$$

and ϕ, Φ are the standard normal probability and cumulative density function respectively. Thus, the observed data likelihood function is not too hard to calculate for this example. It is therefore a good situation to compare MCEM and SSML. Figures 1 and 2 show the results of calculating the MLE for a single 50-observation data set generated from the model given in (8). The exogenous variables x_{i1}, x_{i2} are drawn independently from a uniform distribution on $(-0.5,\, 0.5)$. The true values used for the generation were $\beta = -2$, $\theta = 0.5$.

The MLE for β and θ were found by direct numerical maximization of the likelihood (evaluating it by Laguerra integration, 15 points) to be $\hat{\beta}_{ML} = -0.76$ and $\hat{\theta}_{ML} = 1.01$. This is indicated in Figures 1 and 2 by a dashed line. Three iterative methods were employed to try to reproduce this value: MCEM, SSML (with a simulation sample size of 10 to estimate the likelihood) and SSML (with a simulation sample size of 50 to estimate the likelihood). All three methods used the same starting values. We used the SSML method in conjunction with a GAUSS optimization routine (GAUSS Applications (1992), *Optimization*) to maximize the estimated likelihood. In order to implement MCEM, we selected $nburn = 3 \times iter + 1$, $nskip = 1 + integer(iter/10)$ and $nrep = 5 \times iter + 1$, where $iter$ is the EM iteration. These numbers are small in relation to those recommended in the literature, but we noticed no convergence problems.

Several facts are obvious from the plots and are representative of the general behavior of these estimators.

(i) MCEM approaches a neighborhood of the MLE very quickly, but continues to bounce around the ML estimate even after a fairly long time. The number of Gibbs samples would have to be increased drastically (after approximately 12 iterates) to achieve higher accuracy. Note that this is not done in figures displayed; instead we simply let the simulation sample size increase linearly with the EM iteration. We tried several different starting values and found MCEM to be extremely robust to the choice of starting value.

(ii) The SSML method (using a sample size of 10) is somewhat slower that MCEM. Although it provides an accurate estimate of $\hat{\beta}_{ML}$, it does not yield an accurate estimate of $\hat{\theta}_{ML}$. Even for this simple problem, we found SSML to be sensitive to the choice of starting value. This is not caused by the particular optimization routine used, since the same routine was used within MCEM to find conditional ML estimates of θ with no problems.

(iii) The SSML method (using a sample size of 50) is considerably slower than the other methods, but more accurate. However, the computational time required for this method to converge can increase quite drastically relative to the size of the problem. Sensitivity to starting values was also noticed.

Table 1 compares the performance of these three methods in terms of proximity to the true MLE, after 1 minute, 3 minutes and 15 minutes. The numbers displayed are the average over 50 independent runs of each of these methods. The simulation standard error is reported in parentheses. It is evident that the MCEM iterates approach the MLE much more quickly than the other two methods.

Once the final estimates were obtained, 1000 realizations with 10 cycles between two consecutive realizations were collected to estimate the asymptotic variance-covariance matrix. Table 2 displays the asymptotic variance-covariance matrix of the parameters obtained by Monte Carlo Louis as well as direct numerical integration. A substantial agreement is noted.

Thus, for this simple problem we have demonstrated the feasibility and accuracy of MCEM compared with the SSML method. We also compared the convergence times of MCEM and SSML (using a sample size of 50) for two other simulated data sets. For a three choice problem with three parameters and a data set of size 500, MCEM took approximately 10 minutes to converge while SSML took nearly one hour. For a six choice problem with 15 parameters and a data set of size 1500, the corresponding times were one hour and a day respectively. Thus, while both methods offer a computationally feasible and accurate method of estimation for smaller problems, MCEM continues to be more efficient and just as accurate for larger problems.

5.2 Menu Pricing Data

We now illustrate our MCEM method on data gathered on the quantity demanded of several menu items at a medium-priced family restaurant offering a range of entreés including steak, seafood as well as specialty sandwiches (Kiefer, Kelly and Burdett, 1994). We are interested in studying the demand of four menu items: fish fry, steak and other seafood, Pinesburger and the Tullyburger. Together these items account for more than 65% of the total items ordered. Kiefer, Kelly and Burdett (1994) present a subset of this data on the demand for fish fry, a popular item priced at \$8.95. Of special interest in their study was the effect of changes in price on the demand for fish fry. Four price levels were experimented with (\$8.95, \$9.50, \$9.95, \$10.95). The data were collected over four winter weekends (Friday/Saturday). Actual checks were presented at original price levels so that the customer did not suffer any monetary loss.

Table 3 displays the menu items under study along with the sample frequency of demand over the four weekends, for a total of 974 orders. Our approach is to build a simple choice model to illustrate the computational feasibility of the MCEM method. We re-parametrize the model (as explained in Section 2) using fish fry as the reference item. Nine regressors are considered: three item-specific intercepts, three day-specific intercepts and the effect of price of fish fry on demand. Thus, the latent model for the relative utilities is (dropping the subscript for individual) given by:

$$\begin{aligned} y_{\text{steak}} &= \beta_{\text{steak}} + f\,\gamma_{\text{steak}} + p\,\delta_{\text{steak}} + \epsilon_{\text{steak}} \\ y_{\text{pines}} &= \beta_{\text{pines}} + f\,\gamma_{\text{pines}} + p\,\delta_{\text{pines}} + \epsilon_{\text{pines}} \\ y_{\text{tully}} &= \beta_{\text{tully}} + f\,\gamma_{\text{tully}} + p\,\delta_{\text{tully}} + \epsilon_{\text{tully}} \end{aligned}$$

where the β are item-specific intercepts, f is an indicator for day (1 for Friday and 0 for Saturday),

γ measures the effect of Friday on the different items, and p is the price of fish fry with its corresponding effect δ. The error variance matrix Ω is assumed to be a general 3×3 matrix. Thus, there are 14 identified parameters in this model. Figure 3 graphs the convergence of the MCEM iterates for the fixed effect parameters. Table 4 displays the ML estimates and asymptotic standard errors of these parameters. The Friday effect on the demand of fish fry is clearly seen. The price coefficients are small but not statistically significant. Thus, it appears that even with a fairly large data set, the response of customers to a price change in fish fry is small and insignificant. Our conclusions are consistent with the results of Kiefer, et. al. (1994) who analyzed the data using an independent logit model. Hence a price increase could be supported without serious substitution effects. In fact, the price increase has occurred ($10.95) without any appreciable change in the menu mix or the overall sales.

Figure 4 displays the ML estimates (with standard deviation bars) of the error covariances between menu items. None of the covariances appear to have significant mass away from zero. Thus, the evidence in our data favor an independent probit/logit structure; hence a reduced analysis is adequate.

6 Extensions of MCEM

In this section we describe extensions of our MCEM approach to handle complicated multi-period (panel) choice models as well as non-normal latent distributions.

6.1 Panel Data

Several studies investigate discrete choices made by individuals or households over time (Allenby and Lenk 1992, Geweke, Keane and Runkle, 1994). Such data is known as panel (longitudinal) data. In modeling this data, it is typical to allow the repeated observations on a household to be correlated. We describe how MCEM can be easily adapted to handle panel data, when the relative utilities (for each time period) follow a mixed model. All the ideas extend directly for other general covariance structures (e.g., autoregressive patterns, general covariance patterns, etc.), as well as random coefficient models.

Assume that for each individual i, we observe T_i correlated multinomial vectors $\mathbf{w}_{it} = \{w_{it1}, ..., w_{itc}\}'$.

Further, suppose the latent utilities $\mathbf{y}_{it}$ follow a mixed model in the following manner:

$$\mathbf{y}_{it} \mid \mathbf{b}_i \sim \mathcal{N}_{c-1}(X_{it}\beta + Z_{it}\mathbf{b}_i, \Omega)$$

and

$$\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, D)$$

where Z_{it} is the incidence matrix. Thus, the random effects $\mathbf{b}_i$ provide a convenient mechanism to model the correlation for an individual across time. The parameters of interest are the fixed effects β, the cross correlation parameters in Ω and the serial correlation parameters in D. To implement the MCEM approach, we simply treat the vector of unobserved random effects $\mathbf{b}_i$ as part of the complete data, in addition to the latent utilities $\mathbf{y}_i = \{\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT_i}\}$. A formal description of the algorithm is displayed below:

Step 0) Obtain starting values $\beta^{(0)}$, $\Omega^{(0)}$ and $D^{(0)}$. Set counter $m = 0$.

Step 1) (E-step) For each i and purchase instance t, calculate:

$$E^{(m)}[\mathbf{y}_i \mid \mathbf{w}_i],\ Var^{(m)}[\mathbf{y}_i \mid \mathbf{w}_i],\ E^{(m)}[\mathbf{b}_i \mid \mathbf{w}_i],\text{ and } E^{(m)}[\mathbf{b}_i\,\mathbf{b}_i' \mid \mathbf{w}_i]\,.$$

Step 2a) Set:

$$D^{(m+1)} = N^{-1}\sum_{i=1}^{N} E^{(m)}[\mathbf{b}_i\,\mathbf{b}_i' \mid \mathbf{w}_i]\,.$$

Step 2b) Set:

$$\beta^{(m+1)} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T_i} X_{it}'\,\Omega^{(m)-1}\,X_{it}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T_i} X_{it}'\,\Omega^{(m)-1}\left(E^{(m)}[\mathbf{y}_{it} \mid \mathbf{w}_i] - Z_{it}\,E^{(m)}[\mathbf{b}_i \mid \mathbf{w}_i]\right)\right].$$

Step 2c) Maximize:

$$-\frac{\sum_{i=1}^{N} T_i}{2}\ln|\Omega| - \frac{1}{2}\,tr\left(\Omega^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T_i} Q_{it}(\beta^{(m+1)})\right)$$

over the unknown elements of Ω to obtain $\Omega^{(m+1)}$.

Step 3) If convergence is reached set $\hat{\beta}_{MCEM} = \beta^{(m+1)}$, $\hat{\Omega}_{MCEM} = \Omega^{(m+1)}$ and $\hat{D}_{MCEM} = D^{(m+1)}$; else increment counter m by one and return to Step 1).

In order to estimate the expectations involved in the E-step, we propose the use of the Gibbs sampler to generate from the conditional specifications:

$$[\mathbf{y}_i \mid \mathbf{b}_i, \mathbf{w}_i], \text{ and } [\mathbf{b}_i \mid \mathbf{y}_i, \mathbf{w}_i].$$

This involves generation from truncated normal and normal distributions respectively. Thus, the MCEM implementation is straightforward for this complicated panel data model.

6.2 Non-normal Latent Distributions

A fairly serious criticism of the multinomial probit model is the assumption of multivariate normality for the latent utilities. While this is a mathematically tractable assumption, very little has been done by way of model diagnostics to verify such distributional assumptions. In this section we describe how the MCEM approach does not rely on normality and in fact can quite easily be adapted for non-normal latent distributions. Thus, the MCEM method provides a much more flexible framework of estimation than the simulated likelihood approaches, since the latter are intrinsically dependent on the normality assumption.

Assume that the relative utilities $\mathbf{y}$ are distributed according to some multivariate distribution $g(.)$. In order to implement the Monte Carlo E-step we need to generate variates from the conditional density $[\mathbf{y} \mid \mathbf{w}]$. The advantage of postulating normality is the resulting normality of the full conditional specifications. Hence a Gibbs implementation is straightforward. However, this is not true for any arbitrary distribution $g(.)$. In such cases we propose the use of the Metropolis (Hastings, 1970) algorithm to generate variates from the appropriate conditional densities. The following is an outline of how the Metropolis algorithm is used to generate variates from $[\mathbf{y} \mid \mathbf{w}]$:

Step a) Obtain starting values $\mathbf{y}^{(0)}$. Set counter $m = 0$.

Step b) Generate $\mathbf{y}^* \sim h(.)$ where $h(.)$ is any multivariate density with support on the cone defined by the observed $\mathbf{w}$. A possible choice for $h(.)$ is the truncated normal distribution.

Step c) Calculate the acceptance probability

$$\rho = \min\left(\frac{h(\mathbf{y}^*)\, g(\mathbf{y}^{(m)})}{h(\mathbf{y}^{(m)})\, g(\mathbf{y}^*)}, 1\right).$$

Step d) Accept $\mathbf{y}^{(m+1)}$ as:

$$\mathbf{y}^{(m+1)} = \begin{cases} \mathbf{y}^* & \text{with probability } \rho \\ \mathbf{y}^{(m)} & \text{with probability } (1-\rho). \end{cases}$$

Increment m by one and return to Step b).

Repeat Steps b) through d) a large number of times to form $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, ... Discard a suitable number *nburn* from the beginning of the sequence and then accept every $nskip^{th}$ one to form a sample of size *nrep*. Thus, the ease of implementation of the E-step is unaffected by the choice of latent distribution. However, the M step might be more complicated depending on the choice of $g(.)$.

7 Panel Data Example

This example illustrates the feasibility of MCEM calculations for a simple panel data choice model. We study consumer purchases of six different brands of peanut butter (ERIM panels, A. C. Nielsen, 1985-1988). The six brands under consideration are: Jiff (Creamy & Chunky), Peter Pan (Creamy & Chunky) and Skippy (Creamy & Chunky). Together these brands account for more than 75% of all purchases in the peanut butter category. Our sample of 105 households have 1761 purchase records for these six top brands. Table 5 displays the brands, sample frequency of purchase and the unit price (in cents) averaged over the 1761 records.

The reference brand is Skippy (Creamy). We fit a simple model for the relative utilities:

$$\begin{aligned} \mathbf{y}_{it} &= X_{it}\,\beta + I_5\,\mathbf{b}_i + \epsilon_{it} \\ \mathbf{b}_i &\sim \mathcal{N}_5(\mathbf{0},\, \theta\, I_5) \\ \epsilon_{it} &\sim \mathcal{N}_5(\mathbf{0},\, I_5) \end{aligned}$$

where X_{it} is a 5×6 design matrix corresponding to five brand-specific intercepts and the log of price. The above model allows for household heterogeneity in the intercept. The component of variance, θ captures the variability among household preferences. Table 6 reports the ML estimates (along with asymptotic standard errors) of the parameters β and θ. As might be expected in a choice situation with close substitutes, the price coefficient is large and negative (-6.33). Peter

Pan (Chunky) has the largest negative intercept, indicating that households accord a much lower utility to this brand. This fact is further corroborated by the large negative intercept (-2.00) of the Peter Pan (Creamy) brand as well. The estimate of θ is large and significantly different from zero, supporting the belief that households are extremely heterogeneous in their choice behavior.

8 Discussion

We have developed a flexible framework of estimation which is computationally feasible and accurate for a large class of problems. Our MCEM method overcomes the drawbacks of the simulated likelihood approaches by taking advantage of the latent structure of the MNP model to simplify calculations. For simple problems with a small number of parameters, both methods offer a computationally feasible framework of estimation. However for larger problems, while the simulated likelihood methods become computationally intensive and time-consuming, MCEM continues to offer a flexible and efficient framework of estimation. From our experience we have found MCEM to be robust to choice of starting values; a property that is not shared by the simulated likelihood approaches. Further, MCEM can be very naturally modified to handle more complicated models, like panel data, as well as other models for limited dependent variables (Tobit regression, multinomial logit etc.). Although SSML can accommodate panel data as well, it does so by directly estimating the high-dimensional integral that result from repeated observations on each individual (Supan, Hajivassiliou, Kotlikoff and Morris, 1992). This could exacerbate numerical problems and instabilities, especially since the dimension of these integrals grow very quickly with the number of choices and time periods.

Appendix 1

Hardware and Software Specifications

The random number generator used was a multiplicative-congruential generator (Kennedy, W. J. Jr., and Gentle, J. E. "Statistical Computing", Marcel Dekker, Inc., 1980, pp 136–147). The programs implementing the method proposed in the paper are written in GAUSS (Aptech Systems, 1992). All computations were carried out on Sun (Sparc 10 and 20) workstations.

Appendix 2

DERIVATIVES OF COMPLETE DATA LIKELIHOOD FUNCTION

Expressions for the elements of S and H are:

$$
\begin{aligned}
S_\beta &= \sum_{i=1}^{N} \frac{\partial}{\partial \beta} \ln[\mathbf{y}_i \mid \beta,\, \Omega] \\
&= \sum_{i=1}^{N} X_i' \, \Omega^{-1} \, \mathbf{e}_i \\
S_{\omega_{rs}} &= \sum_{i=1}^{N} \frac{\partial}{\partial \omega_{rs}} \ln[\mathbf{y}_i \mid \beta,\, \Omega] \\
&= -\frac{N}{2}\, tr(\Omega^{-1}\, \dot{\Omega}_{rs}) + \frac{1}{2} \sum_{i=1}^{N} \mathbf{e}_i'\Omega^{-1}\dot{\Omega}_{rs}\Omega^{-1}\mathbf{e}_i \\
H_{\beta\beta} &= \sum_{i=1}^{N} \frac{\partial^2}{\partial \beta^2} \ln[\mathbf{y}_i \mid \beta,\, \Omega] \\
&= -\sum_{i=1}^{N} X_i'\Omega^{-1}X_i \\
H_{\beta\omega_{rs}} &= \sum_{i=1}^{N} \frac{\partial^2}{\partial \beta\, \partial \omega_{rs}} \ln[\mathbf{y}_i \mid \beta,\, \Omega] \\
&= -\sum_{i=1}^{N} X_i'\Omega^{-1}\dot{\Omega}_{rs}\Omega^{-1}\mathbf{e}_i \\
H_{\omega_{rs}\omega_{tu}} &= \sum_{i=1}^{N} \frac{\partial^2}{\partial \omega_{tu}\, \partial \omega_{rs}} \ln[\mathbf{y}_i \mid \beta,\, \Omega] \\
&= \frac{N}{2}\, tr(\Omega^{-1}\dot{\Omega}_{tu}\Omega^{-1}\dot{\Omega}_{rs}) - \sum_{i=1}^{N} \mathbf{e}_i'\Omega^{-1}(\dot{\Omega}_{tu}\Omega^{-1}\dot{\Omega}_{rs} + \dot{\Omega}_{rs}\Omega^{-1}\dot{\Omega}_{tu})\Omega^{-1}\mathbf{e}_i
\end{aligned}
$$

where ω_{rs} is the $(r,s)^{th}$ element of Ω, $\dot{\Omega}_{rs} = \frac{\partial}{\partial \omega_{rs}}\Omega$ and $\mathbf{e}_i = (\mathbf{y}_i - X_i\beta)$.

References

ALLENBY, G. AND LENK, P. (1992) Modeling Household Purchase Behavior, *Series Paper WPS 92-39*, Business Research, College of Business, The Ohio State University.

APTECH SYSTEMS (1992) The GAUSS System Version 3.0, Maple Valley, Washington 98038.

ASHFORD, J. R., AND SOWDEN, R. R. (1970) Multi-Variate Probit Analysis, *Biometrics* **26:** 535–546.

BÖRSCH-SUPAN, A., HAJIVASSILIOU, V., KOTLIKOFF, L. J., AND MORRIS, J. N. (1992) Health, Children, and Elderly Living Arrangements: A Multi-period-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).

BÖRSCH-SUPAN, A. AND HAJIVASSILIOU, V. (1993) Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models, *Journal of Econometrics* **58:** 347–368.

CHAN, K. S., AND LEDOLTER, J. (1995) Monte Carlo EM Estimation for Time Series Models Involving Counts, *Journal of the American Statistical Association* **90:** 242–252.

DAGANZO, C. (1980) *Multinomial Probit*, New York, Acedemic Press.

DANSIE, B. 1985 Parameter Estimability in the Multinomial Probit Model, *Transportation Research* **19B:** 526–528.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977) Maximum Likelihood Estimation From Incomplete Observations, *Journal of the Royal Statistical Society, Ser B* **39:** 1–38.

DEVROYE, L. (1986) *Non-Uniform Random Variate Generation*, Springer-Verlag, New-York.

GEMAN, S., AND GEMAN, D. (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6:** 721–741.

GEWEKE, J. F. (1989) Efficient Simulation from the Multivariate Normal Distribution Subject to Linear Inequality Constraints and the Evaluation of Constraint Probabilities, *Mimeo* Duke University, Durham, NC.

GEWEKE, J. F., KEANE, M. P. AND RUNKLE, D. E. (1994) Statistical Inference in the Multinomial Multiperiod Probit Model, *Research Department Staff Report 177*, Federal Reserve Bank of Minneapolis.

GUO, S. W. AND THOMPSON, E. A. (1992) A Monte Carlo Method for Combined Segregation and Linkage Analysis, *American Journal of Human Genetics* **51:** 1111–1126.

HAJIVASSILIOU, V. AND MCFADDEN, D. (1990) The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crisis, Working Paper (Yale University, New Haven, CT).

HAJIVASSILIOU, V. AND MCFADDEN, D. RUUD (1992) Simulation of multivariate normal orthant probabilities: Methods and Programs. Cowles Foundation Discussion Paper 1021. Yale University.

HASTINGS, W. K. (1970) Monte Carlo Sampling Methods using Markov Chains and their Applications, *Biometrika* **57:** 97–109.

HAUSMAN, J. AND WISE D. (1978) A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences, *Econometrica* **45:** 319–339.

JENNRICH, R. I. AND SCHLUCHTER, M. D. (1986) Unbalanced Repeated Measures Models with Structured Covariance Matrices, *Biometrics* **42:** 805–820.

KEANE, M. (1993) A Computationally Practical Simulation Estimator for Panel Data, *Econometrica* forthcoming.

KIEFER, N. M., KELLY, T. J., AND BURDETT, K. (1994) Menu Pricing: An Experimental Approach, *Journal of Business and Economic Statistics* Vol 12 **3:** 329–337.

LERMAN, S. AND MANSKI, C. (1981) On the Use of Simulated Frequencies to Approximate Choice Probabilities, in: C. Manski and D. McFadden eds., *Structural Analysis of Discrete Data with Econometric Applications* (MIT Press, Cambridge MA) 305–319.

LOUIS, T. A. (1982) Finding the Observed Information Matrix when Using the EM Algorithm, *Journal of the Royal Statistical Society, Ser B* **44:** 226–233.

MCCULLOCH, C. E. (1993) Maximum Likelihood Variance Components Estimation for Binary Data, *Journal of the American Statistical Association* **89:** 330–335.

MCCULLOCH, R. AND ROSSI, P. (1994) An Exact Likelihood Analysis of the Multinomial Probit Model, *Journal of Econometrics* **64:** 207–240.

MCFADDEN, D. (1989) A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration, *Econometrica* **57:** 995–1027.

MENG, X. L. AND RUBIN, D. B. (1993) Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework, *Biometrika* **80:** 267–278.

NATARAJAN, R. AND MCCULLOCH, C. E. (1995) A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses, *Biometrika* **82:** forthcoming.

SEARLE, S., CASELLA, G., AND MCCULLOCH, C. E. (1992) *Variance Components*, Wiley, NY.

STERN, S. (1992) A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models, *Econometrica* **60:** 943–952.

TANNER, M. A. (1993) *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions* Second Edition, Springer-Verlag.

WEI, G. C. G, AND TANNER, M. A. (1990) A Monte Carlo Implementation of the EM algorithm and the Poor Man's Data Augmentation Algorithm. *Journal of the American Statistical Association* **85**, 829–839.

CAPTIONS FOR TABLES

Table 1: MCEM and SSML Iterates after 1 minute, 3 minutes and 15 minutes. (Averaged over 50 independent runs; Simulation standard errors reported in parentheses).

Table 2: Asymptotic Variance-Covariance Matrix of ML Estimates Obtained by Numerical Integration and Monte Carlo Louis.

Table 3: Purchase Frequencies of Items from the Menu-Pricing Data.

Table 4: Parameter Estimates (Asymptotic Standard Error) for 4-Choice Model for the Menu-Pricing Data.

Table 5: Purchase Frequencies and Costs for the Peanut Butter Data.

Table 6: Parameter Estimates (Asymptotic Standard Error) for 6-Choice Hierarchical Model for the Peanut Butter Data.

CAPTIONS FOR FIGURES

Figure 1: Convergence of MCEM and SSML for β.

Figure 2: Convergence of MCEM and SSML for θ.

Figure 3: Convergence of MCEM Iterates for Fixed Effects Parameters for Menu-Pricing Data.

Figure 4: ML Estimates (Standard Error Bars) of Error Covariances.

Parameter	True MLE	Elapsed CPU Time (in minutes)	Parameter Estimate MCEM	SSML (10)	SSML (50)
β	-0.76	1	-0.7534	-0.7399	-0.7350
			(0.0074)	(0.0067)	(0.0303)
		3	-0.7599	-0.7591	-0.8702
			(0.0065)	(0.0057)	(0.0053)
		15	-0.7601	-0.7591	-0.7581
			(0.0053)	(0.0057)	(0.0033)
θ	1.01	1	0.9577	1.1970	0.7346
			(0.0359)	(0.0422)	(0.0046)
		3	1.0375	1.1043	1.0851
			(0.0248)	(0.0275)	(0.0122)
		15	1.0381	1.1043	1.0095
			(0.0202)	(0.0275)	(0.0146)

Monte Carlo Louis			Numerical Integration		
	β	θ		β	θ
β	0.3945	-0.2705	β	0.3966	-0.2735
θ	-0.2705	2.1254	θ	-0.2735	2.1538

Menu Item	Frequency of Demand
Steak and Other Seafood	0.22
Fish Fry	0.22
Pinesburger	0.43
Tullyburger	0.13

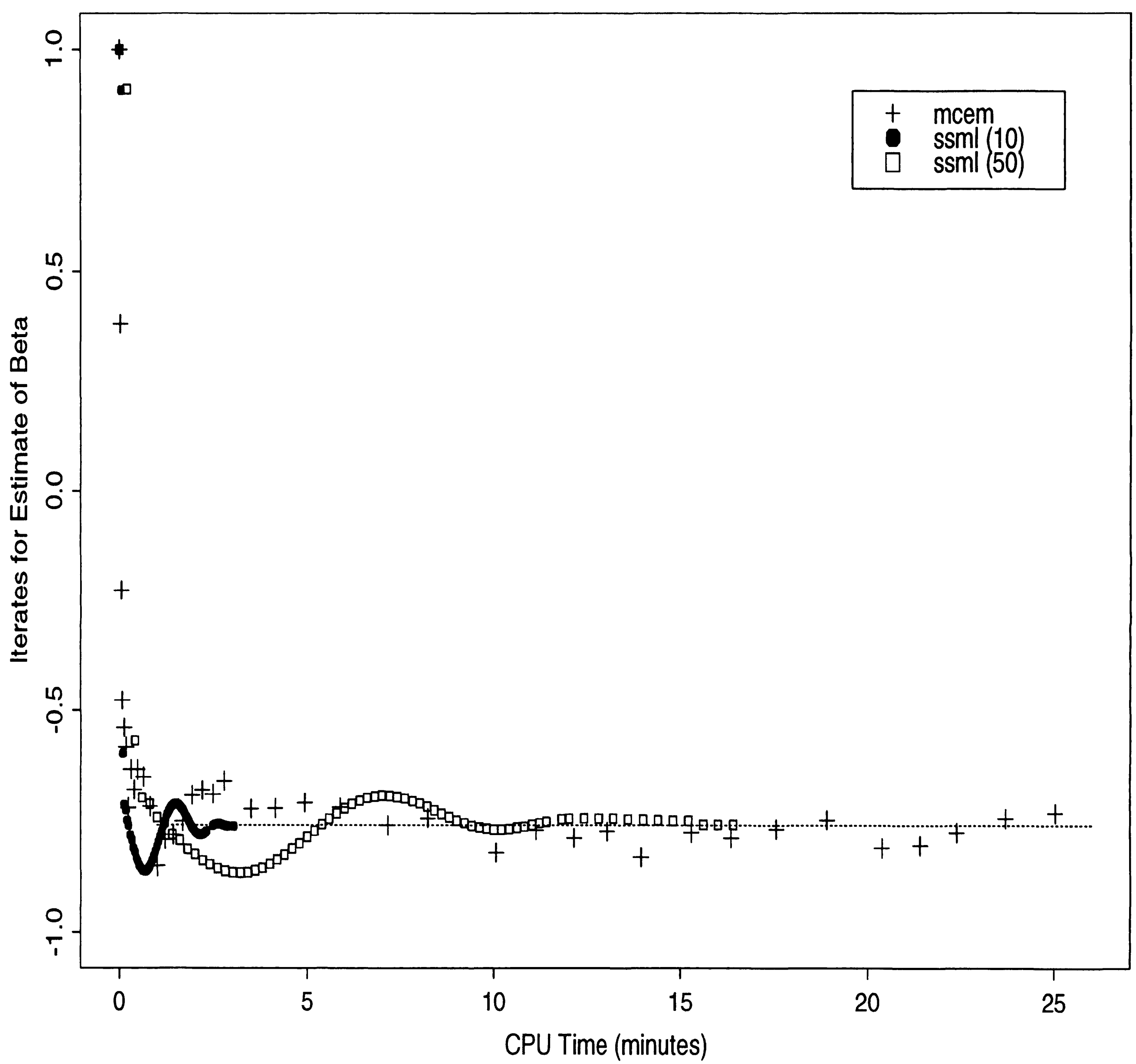
	Parameter Estimate	Asymptotic Std Error
Item Intercept		
β_{steak}	0.1683	0.3898
β_{pines}	0.7450*	0.2606
β_{tully}	-0.3619	0.3400
"Friday Effect"		
γ_{steak}	-0.5016*	0.1781
γ_{pines}	-0.7970*	0.1703
γ_{tully}	-0.3133*	0.1563
Price Coefficient		
δ_{steak}	-0.0719	0.0716
δ_{pines}	-0.0511	0.0703
δ_{tully}	0.0950	0.0762

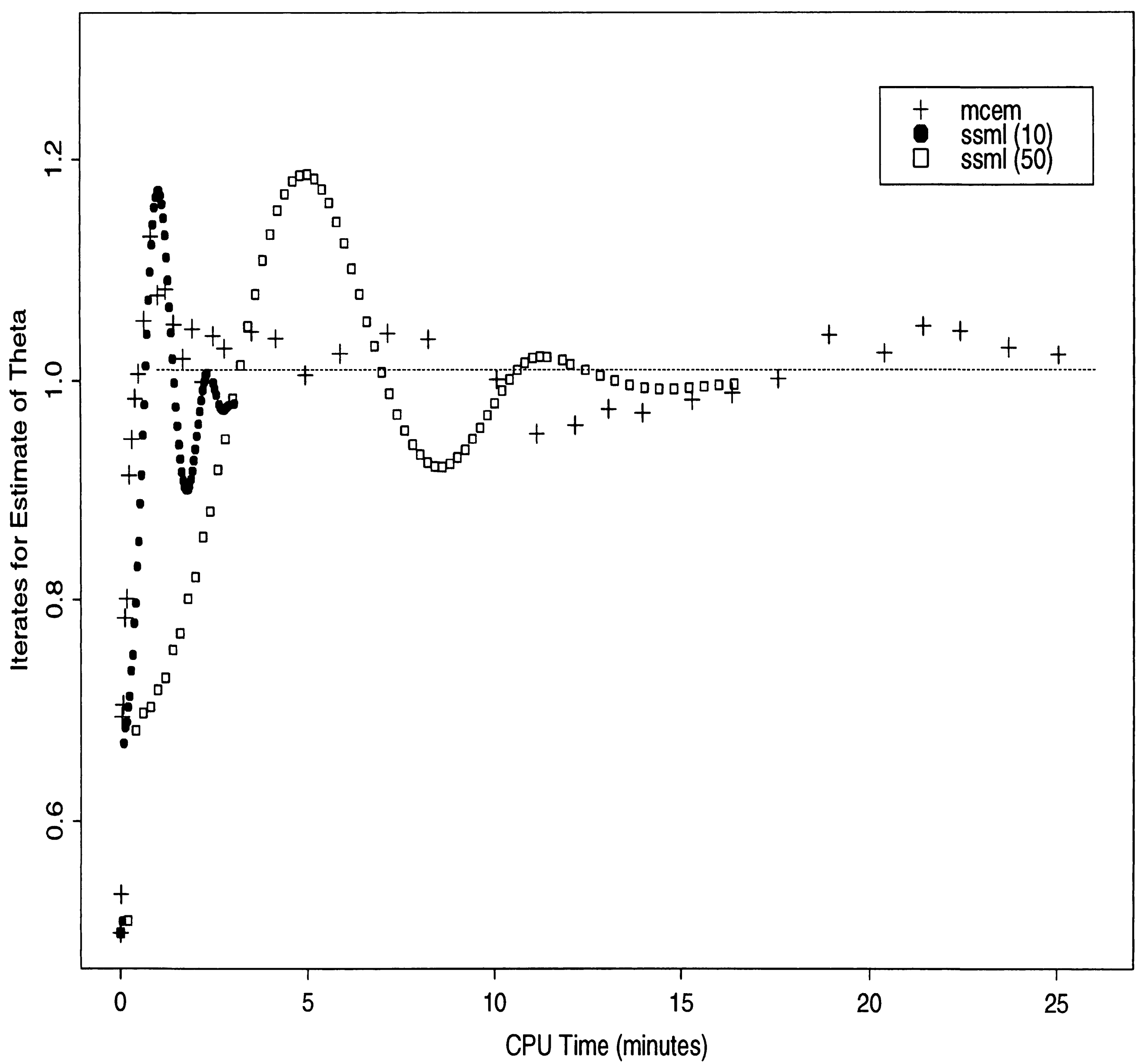
* : statistically significantly different from 0 at the 5% level.

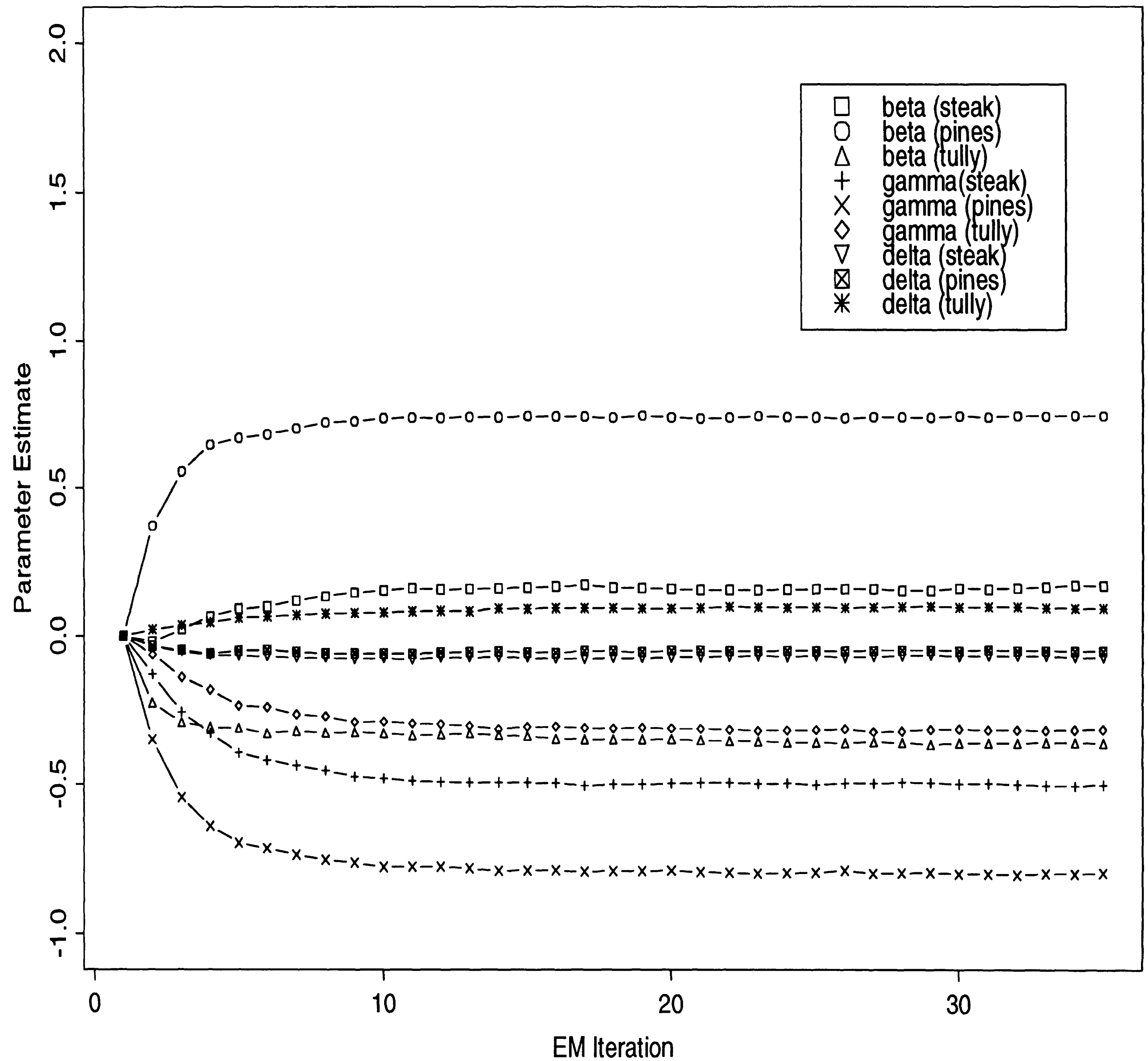
Brand	Frequency of Purchase	Unit Price (in cents)
Jiff (Creamy)	0.26	9.81
Jiff (Chunky)	0.08	10.00
Peter Pan (Creamy)	0.07	9.61
Peter Pan (Chunky)	0.03	9.96
Skippy (Creamy)	0.33	9.32
Skippy (Chunky)	0.22	9.18

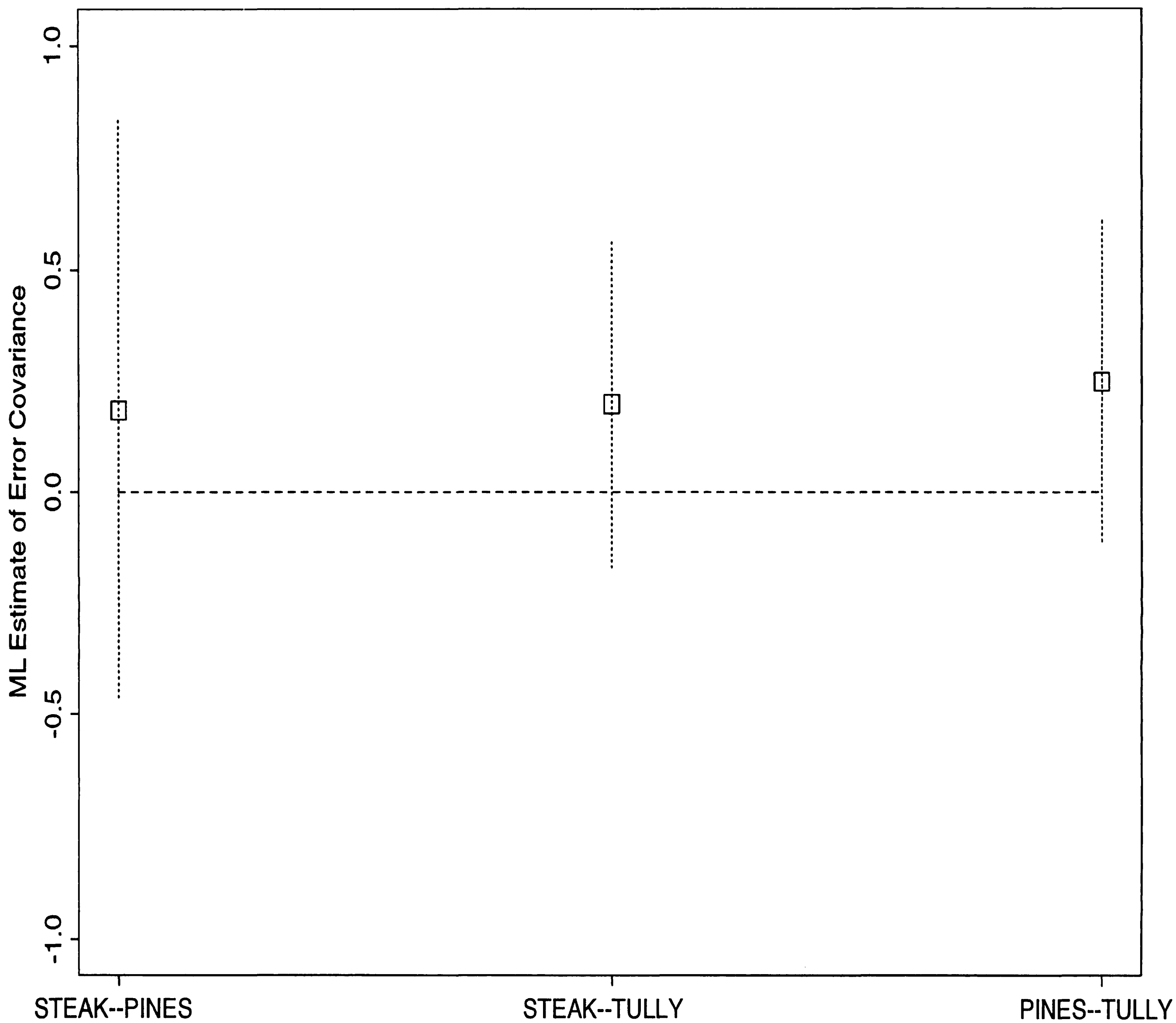
	Parameter Estimate	Asymptotic Std Error
Brand Intercept		
Jiff (Creamy)	-0.7212*	0.1743
Jiff (Chunky)	-1.7177*	0.1930
Peter Pan (Creamy)	-1.9813*	0.1972
Peter Pan (Chunky)	-2.4235*	0.2099
Skippy (Chunky)	-0.9683*	0.1725
Price Coefficient	-6.3353*	0.5946
Variance Component	2.4510*	0.0714

* : statistically significantly different from 0 at the 5% level.


mcem
ssml (10)
ssml (50)
Iterates for Estimate of Beta
CPU Time (minutes)
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
25


mcem
ssml (10)
ssml (50)
1.2
1.0
0.8
0.6
Iterates for Estimate of Theta
0
5
10
15
20
25
CPU Time (minutes)


beta (steak)
beta (pines)
beta (tully)
gamma(steak)
gamma (pines)
gamma (tully)
delta (steak)
delta (pines)
delta (tully)
Parameter Estimate
EM Iteration
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
0
10
20
30


ML Estimate of Error Covariance
1.0
0.5
0.0
-0.5
-1.0
STEAK--PINES
STEAK--TULLY
PINES--TULLY