A Recursive Estimator for Random Coefficient Models

Kenneth Train
Department of Economics
University of California, Berkeley

October 18, 2007

Abstract

This paper describes a recursive method for estimating random coefficient models. Starting with a trial value for the moments of the distribution of coefficients in the population, draws are taken and then weighted to represent draws from the conditional distribution for each sampled agent (i.e., conditional on the agent's observed dependent variable.) The moments of the weighted draws are calculated and then used as the new trial values, repeating the process to convergence. The recursion is a simulated EM algorithm that provides a method of simulated scores estimator. The estimator is asymptotically equivalent to the maximum likelihood estimator under specified conditions. The recursive procedure is faster than maximum simulated likelihood (MSL) with numerical gradients, easier to code than MSL with analytic gradients, assures a positive definite covariance matrix for the coefficients at each iteration, and avoids the numerical difficulties that often occur with gradient-based optimization. The method is illustrated with a mixed logit model of households' choice among energy suppliers.

Keywords: Mixed logit, probit, random coefficients, EM algorithm.

1 Introduction

Random coefficient models, such as mixed logit or probit, are widely used because they parsimoniously represent the fact that different agents have different preferences. The parameters of the model are the parameters of the distribution of coefficients in the population. The specifications generally permit full covariance among the random coefficients. However, this full generality is seldom realized in empirical applications due to the numerical difficulty of maximizing a likelihood function that contains so many parameters. As a result, most applications tend to assume no covariance among coefficients (Chen and Cosslett, 1998, Goett et al., 2000, Hensher et al., 2005) or covariance among only a subset of coefficients (Train, 1998, Revelt and Train, 1998).[1]

This paper presents a procedure that facilitates estimation of random coefficient models with full covariance among coefficients. In its simplest form, it is implemented as follows. For each sampled agent, draws are taken from the population distribution of coefficients using a trial value for the mean and covariance of this distribution. Each draw is weighted proportionally to the probability of the agent's observed dependent variable under this draw. The mean and covariance of these weighted draws over all sampled agents are then calculated. This mean and covariance become the new trial values, and the process is repeated to convergence. The procedure provides a method of simulated scores estimator (Hajivassiliou and McFadden, 1998), which is asymptotically equivalent to maximum likelihood under well-known conditions discussed below. The recursive procedure constitutes a simulated EM algorithm (Dempster et al., 1977; Ruud, 1991), which converges to a root of the score condition.

The procedure is related to the diagnostic tool described by Train (2003, section 11.5) of comparing the conditional and unconditional densities of co-

[1]Restrictions on the covariances are not as benign as they might at first appear. For example, Louviere (2003) argues, with compelling empirical evidence, that the scale of utility (or, equivalently, the variance of random terms over repeated choices by the same agent) varies over people, especially in stated-preference experiments. Models without full covariance of utility coefficients imply the same scale for all people. If in fact scale varies, the variation in scale, which does not affect marginal rates of substitution (MRS), manifests erroneously as variation in independent coefficients that does affect estimated MRS.

efficients for an estimated model. In particular, to evaluate a model, draws are taken from the conditional distribution of coefficients for each agent in the sample, and then the distribution of these draws is compared with the estimated population (i.e., unconditional) distribution. If the model is correctly specified, the two distributions should be similar, since the expectation of the former is equal to the later. In the current paper, this concept is used as an estimation criterion rather than a diagnostic tool.

The procedure is described and applied in the sections below. Section 2 provides the basic version under assumptions that are more restrictive than needed but facilitate explanation and implementation. Section 3 generalizes the basic version. Section 4 applies the procedure to data on households' choices among energy suppliers.

2 Basic Version

Each agent faces exogenous observed explanatory variables x and observed dependent variable(s) y. We assume in our notation that y is discrete and x is continuous, though these assumptions can be changed with appropriate change in notation. Let β be a vector of random coefficients that affect the agent's outcome and are distributed over agents in the population with density $f(\beta \mid \theta)$, where θ are parameters that characterize the density, such as its mean and covariance. For the purposes of this section, we specify f to be the normal density, independent of x; these assumptions will be relaxed in section 3. Let $m(\beta)$ be the vector-valued function consisting of β itself and the vectorized lower portion of $(\beta\beta')$, Then, by definition, $\theta = \int m(\beta) f(\beta \mid \theta) d\beta$. That is, θ are the unconditional moments of β.

Consider now the behavioral model. Given β, the behavioral model gives the probability that an agent facing x has outcome y as some function $L(y \mid \beta, x)$, which we assume in this section depends on coefficients β and not (directly) on elements of θ. In a mixed logit model with repeated choices for each agent, L is a product of logits. In other models, L, which we call the kernel of the behavioral model, takes other forms.[2] Since β is not known, the probability

[2]If all random elements of the behavioral model are captured in β, then L is an indicator

of outcome y is $P(y \mid x, \theta) = \int L(y \mid \beta, x) f(\beta \mid \theta) d\beta$.

The density of β can be determined for each agent conditional on the agent's outcome. This conditional distribution is the distribution of β among the subpopulation of agents who, when faced with x, have outcome y. By Bayes' identity, the conditional density is $h(\beta \mid y, x, \theta) = L(y \mid \beta, x) f(\beta \mid \theta) / P(y \mid x, \theta)$. The moments of this conditional density are $\int m(\beta) h(\beta \mid y, x, \theta) d\beta$, and the expectation of such moments in the population is:

$$M(\theta) = \int_x \sum_y \mathcal{S}(y \mid x) \int_\beta m(\beta) h(\beta \mid y, x, \theta) \, d\beta \, g(x) \, dx$$

where $g(x)$ is the density of x in the population and $\mathcal{S}(y \mid x)$ is the share of agents with outcome y among those facing x.

Denote the true parameters as θ^*. At the true parameters $\mathcal{S}(y \mid x) = P(y \mid x, \theta^*)$, such that the expected value of the moments of the conditional distributions equals the unconditional moments:

$$\begin{aligned}
M(\theta^*) &= \int_x \sum_y \mathcal{S}(y \mid x) \int_\beta m(\beta) \frac{L(y \mid \beta, x) f(\beta \mid \theta^*) d\beta}{P(y \mid x, \theta^*)} g(x) dx \\
&= \int_x \sum_y \int_\beta m(\beta) L(y \mid \beta, x) f(\beta \mid \theta^*) d\beta g(x) dx \\
&= \int_x \int_\beta m(\beta) [\sum_y L(y \mid \beta, x)] f(\beta \mid \theta^*) d\beta g(x) dx \\
&= \int_x \int_\beta m(\beta) f(\beta \mid \theta^*) d\beta g(x) dx \\
&= \theta^*.
\end{aligned}$$

since $L(y \mid \beta, x)$ sums to one over all possible values of y.

The estimation procedure uses a sample analog to the population expectation $M(\theta)$. The variables for sampled agents are subscripted by $n = 1, ..., N$. The sample average of the moments of the conditional distributions is then:

$$\mathcal{M}(\theta) = \frac{1}{N} \sum_n \int_\beta m(\beta) \frac{L(y_n \mid \beta, x_n)}{P(y_n \mid x_n, \theta)} f(\beta \mid \theta) d\beta.$$

This quantity is simulated as follows: (1) For each agent, take R draws of β from $f(\beta \mid \theta)$ and label the r-th draw for agent n as β_{nr}. (2) Calculate $L(y_n \mid \beta_{nr}, x_n)$

function of whether or not the observed outcome arises under that β.

for all draws for all agents. (3) Weight draw β_{nr} by $w_{nr} = \frac{L(y_n|\beta_{nr},x_n)}{\frac{1}{R}\sum_{r'} L(y_n|\beta_{nr'},x_n)}$, such that the weights average to one over draws for each given agent. (4) Average the weighted moments:

$$\tilde{\mathcal{M}}(\theta) = \sum_n \sum_r w_{nr}\, m(\beta_{nr})/NR$$

The estimator $\hat{\theta}$ is defined by $\tilde{\mathcal{M}}(\hat{\theta}) = \hat{\theta}$. The recursion starts with an initial value of θ and repeatedly calculates $\theta_{t+1} = \tilde{\mathcal{M}}(\theta_t)$ until $\theta_{T+1} = \theta_T$ within a tolerance. Since the first two moments determine the covariance, the procedure is equivalently applied to the mean and covariance directly. Note that the covariance in each iteration is necessarily positive definite, since it is calculated as the covariance of weighted draws.

We first examine the properties of the estimator and then the recursion.

2.1 Relation of estimator to maximum likelihood

Given the specification of $P(y_n \mid x_n, \theta)$, the score can be written:

$$\begin{aligned} s_n(\theta) &= \frac{\partial log P(y_n \mid x_n, \theta)}{\partial \theta} \\ &= \frac{1}{P(y_n \mid x_n, \theta)} \int L(y_n \mid \beta, x_n) \frac{\partial f(\beta \mid \theta)}{\partial \theta} d\beta \\ &= \int \frac{\partial log f(\beta \mid \theta)}{\partial \theta} \frac{L(y_n \mid \beta, x_n)}{P(y_n \mid x_n, \theta)} f(\beta \mid \theta) d\beta. \end{aligned}$$

The maximum likelihood estimator is a root of $\sum_n s_n(\theta) = 0$.

Let b be the mean and W the covariance of the normally distributed coefficients, such that $log f(\beta \mid b, W) = k - \frac{1}{2} log(|W|) - \frac{1}{2}(\beta - b)' W^{-1}(\beta - b)$. The derivatives entering the score are:

$$\frac{\partial log f}{\partial b} = -W^{-1}(\beta - b)$$

$$\frac{\partial log f}{\partial W} = -\frac{1}{2} W^{-1} + \frac{1}{2} W^{-1}[(\beta - b)(\beta - b)']W^{-1}.$$

It is easy to see that $\sum_n s_n(\theta^0) = 0$ for some θ^0 if and only if $\mathcal{M}(\theta^0) = \theta^0$, such that, in the non-simulated version, the estimator is the same as MLE.

Consider now simulation. A direct simulator of the score is

$$\tilde{s}_n(\theta) = \frac{1}{R} \sum_r w_{nr} \frac{\partial log f(\beta \mid \theta)}{\partial \theta}.$$

A method of simulated scores estimator is a root of $\sum_n \tilde{s}_n(\theta) = 0$. As in the non-simulated case, $\sum \tilde{s}_n(\theta^0) = 0$ iff $\tilde{\mathcal{M}}(\theta^0) = \theta^0$, such that the recursive estimator is this MSS estimator. Hajivassiliou and McFadden (1998) give properties of MSS estimators. In our case, the score simulator is not unbiased, due to the inverse probability that enters the weights. In this case, the MSS estimator is consistent and asymptotically equivalent to MLE if R rises at a rate greater than $\sqrt{N}$.

These properties, and the requirement on the draws, are the same as for maximum simulated likelihood (MSL; Hajivassiliou and Ruud, 1994, Lee, 1995.) However, the estimator is not the same as the MSL estimator. For MSL, the probability is expressed as an integral over a parameter-free density, with the parameters entering the kernel. The gradient then involves the derivatives of the kernel rather than the derivatives of the density. That is, the coefficients are treated as functions $\beta(\theta, \mu)$ with μ having a parameter-free distribution. The probability is expressed as $P(y \mid x, \theta) = \int L(y \mid \beta(\theta, \mu), x) f(\mu) d\mu$ and simulated as $\tilde{P}(y \mid x, \theta) = \sum_r L(y \mid \beta(\theta, \mu_r), x)/R$ for draws $\mu_1, ...\mu_R$. The derivative of the log of this simulated probability is

$$\tilde{\tilde{s}}(\theta) = \frac{1}{\tilde{P}(y \mid x, \theta)} \frac{1}{R} \sum_r \frac{\partial L(y \mid \beta(\theta, \mu_r), x)}{\partial \theta},$$

which is not numerically the same as $\tilde{s}(\theta)$ for a finite number of draws. In particular, the value of θ that solves $\sum_n \tilde{s}_n(\theta) = 0$ is not the same as the value that solves $\sum_n \tilde{\tilde{s}}_n(\theta) = 0$ and maximizes the log of the simulated likelihood function. Either simulated score can serve as the basis for a MSS estimator, and they are asymptotically equivalent to each other under the maintained condition that R rises faster than $\sqrt{N}$. The distinction is the same as for any MSS estimator that is based on a simulated score that is not the derivative of the log of the simulated probability.[3]

The simulated scores at $\hat{\theta}$ provide an estimate of the information matrix, analogous to the BHHH estimate for standard maximum likelihood: $\hat{\mathcal{I}} = S'S/N$, where S is the NxK matrix of K-dimensional scores for N agents.

[3]An important class are the unbiased score simulators that Hajivassiliou and McFadden (1998) discuss, which, by definition, differ from the derivative of the log of the simulated probability because the latter is necessarily biased due to the log operation.

The covariance matrix of the estimated parameters is then estimated as $V = \hat{\mathcal{I}}^{-1}/N = (S'S)^{-1}$, under the maintained assumption that R rises faster than $\sqrt{N}$. Also, the scores can be used as a convergence criterion, using the statistic $\bar{s}'V\bar{s}$, where $\bar{s} = \sum_n \tilde{s}_n/N$.

2.2 Simulated EM algorithm

We can show that the recursive procedure is an EM algorithm and, as such, is guaranteed to converge. In general, an EM algorithm is a procedure for maximizing a likelihood function in the presence of missing data (Dempster, et al., 1977). For sample $n = 1, \ldots, N$, with discrete observed sample outcome y_n and continuous missing data z_n for observation n (and suppressing notation for observed explanatory variables), the likelihood function is $\sum_n log \int P(y_n \mid z, \theta) f_n(z \mid \theta) dz$, where $f_n(z \mid \theta)$ is the density of the missing data for observation n which can depend on parameters θ. The recursion is specified as:

$$\theta_{t+1} = argmax_\theta \sum_n \int h_n(z \mid y_n, \theta_t) logP(y_n, z \mid \theta) dz$$

where P is the probability-density of both the observed outcome and missing data, and h is the density of the missing data conditional on y. It is called EM because it consists of an expectation that is maximized. The term being maximized is the expected log-likelihood of both the outcome and the missing data, where this expectation is over the density of the missing data conditional on the outcome. The expectation is calculated using the previous iteration's value of θ in h_n, and the maximization to obtain the next iteration's value is over θ in $logP(y_n, z \mid \theta)$. This distinction between the θ entering the weights for the expectation and the θ entering the log-likelihood is the key element of the EM algorithm. Under conditions given by Boyles (1983) and Wu (1983), this algorithm converges to a local maximum of the original likelihood function. As with standard gradient-based methods, it is advisable to check whether the local maximum is global, by, e.g., using different starting values.

In the present context, the missing data are the β's which have the same unconditional density for all observations, such that the above notation is trans-

lated to z_n=β_n and $f_n(z \mid \theta) = f(\beta \mid \theta) \forall n$. The EM recursion becomes:

$$\theta_{t+1} = argmax_\theta \sum_n \int h(\beta \mid y_n, x_n, \theta_t) log[L(y_n \mid \beta, x_n) f(\beta \mid \theta)] d\beta. \quad (1)$$

Since $L(y_n \mid \beta, x_n)$ does not depend on θ, the recursion becomes

$$\theta_{t+1} = argmax_\theta \sum_n \int h(\beta \mid y_n, x_n, \theta_t) log f(\beta \mid \theta) d\beta \quad (2)$$

The integral is approximated by simulation, giving:

$$\theta_{t+1} = argmax_\theta \sum_n \sum_r w_{nr}(\theta_t) log f(\beta_{nr} \mid \theta) \quad (3)$$

where the weights are expressed as functions of θ_t since they are calculated from θ_t. Note, as stated above, that in the maximization to obtain θ_{t+1}, the weights are fixed, and the maximization is over θ in f. The function being maximized is the log-likelihood function for a sample of draws from f weighted by $w(\theta_t)$. In the current section, f is the normal density, which makes this maximization easy. In particular, for a sample of weighted draws from a normal distribution, the maximum likelihood estimator for the mean and covariance of the distribution is simply the mean and covariance of the weighted draws. This is our recursive procedure.[4]

3 Generalization

We consider non-normal distributions, fixed coefficients, and parameters that enter the kernel but not the distribution of coefficients.

3.1 Non-normal distributions

For distributions that can be expressed as a transformation of normally distributed terms, the transformation can be taken in the kernel, $L(y \mid T(\beta), x)$

[4]EM algorithms have been used extensively to examine Gaussian mixture models for cluster analysis and data mining (e.g., McLachlan and Peel, 2000.) In these models, the density of the data is described by a mixture of Gaussian distributions, and the goal is to estimate the mean and covariance of each Gaussian distribution and the parameters that mix them. In our case, the data being explained are discrete outcomes rather than continuous variables, and the Gaussian is the mixing distribution rather than the quantity that is mixed.

for transformation T, and all other aspects of the procedure remain the same. The parameters of the model are still the mean and covariance of the normally distributed terms, before transformation. Draws are taken from a normal with given mean and covariance, weights are calculated for each draw, the mean and covariance of the weighted draws are calculated, and the process is repeated with the new mean and covariance. The transformation affects the weights, but nothing else. A considerable degree of flexibility can be obtained in this way. Examples include lognormals with transformation $exp(\beta)$, censored normal with $max(0, \beta)$, and Johnson's S_B distribution with $exp(\beta)/(1 + exp(\beta))$. The empirical application in section 4 explores the use of these kinds of transformations.

For any distribution, the EM algorithm in eqn (4) states that the next value of the parameter, θ_{t+1}, is the MLE of θ from a sample of weighted draws from the distribution. With a normal distribution, the MLE is the mean and covariance of the weighted draws. For many other distributions, the same is true, namely, that the parameters of the distribution are moments whose MLE is the analogous moments in the sample of weighted draws. When this is not the case, then the moments of the weighted draws are replaced with whatever constitutes the MLE of parameters based on the weighted draws. The equivalence of $\sum \tilde{s}_n(\theta) = 0$ and $\tilde{\mathcal{M}}(\theta) = \theta$ arises under any f when $\tilde{\mathcal{M}}$ is defined as the MLE estimator from weighted draws from f.

3.2 Fixed coefficients and parameters in the kernel

The procedure can be conveniently modified to allow random coefficients to contain a systematic part that would ordinarily appear as a fixed coefficient in the kernel. Let $\beta_n = \Gamma z_n + \eta_n$ where z_n is a vector of observed variables relating to agent n, Γ is a conforming matrix, and η_n is normally distributed. The parameters θ are now Γ and the mean and covariance of η. The density of β is denoted $f(\beta \mid z_n, \theta)$ since it depends on z. The probability for observation n is $P(y_n \mid x_n, z_n, \theta) = \int L(y_n \mid \beta, x_n) f(\beta \mid z_n, \theta) d\beta$, and the conditional density of β is $h(\beta \mid y_n, x_n, z_n, \theta_t) = L(y_n \mid \beta, x_n) f(\beta \mid z_n, \theta) / P(y_n \mid x_n, z_n, \theta)$. The EM

recursion is

$$\theta_{t+1} = argmax_\theta \sum_n \int h(\beta \mid y_n, x_n, z_n, \theta_t) log[L(y_n \mid \beta, x_n) f(\beta \mid z_n, \theta)] d\beta.$$

As before, L does not depend on θ and so drops out, giving:

$$\theta_{t+1} = argmax_\theta \sum_n \int h(\beta \mid y_n, x_n, z_n \theta_t) log f(\beta \mid z_n, \theta) d\beta.$$

which is simulated by

$$\theta_{t+1} = argmax_\theta \sum_n \sum_r w_{nr}(\theta_t) log f(\beta_{nr} \mid z_n, \theta) \tag{4}$$

where $w_{nr} = L(y_n \mid \beta_{nr}, x_n) / \sum_{r'} \frac{1}{R} L(y_n \mid \beta_{nr'}, x_n)$. Given a value of θ, draws of β_n are obtained by drawing η from its normal distribution and adding Γz_n. The weight for each draw of β_n is determined as before, proportional to $L(y_n \mid \beta_n, x)$. Then the ML estimate of θ is obtained from the sample of weighted draws. Since β is specified as a system of linear equations with normal errors, the MLE of the parameters is the weighted seemingly unrelated regression (SUR) of β_n on z_n (e.g., Greene, 2000, section 15.4). The estimated coefficients of z_n are the new value of Γ; the estimated constants are the new means of η; and the covariance of the residuals is the new value of the covariance of η.

For fixed parameters that are not implicitly part of a random coefficient, an extra step must be added to the procedure. To account for this generality, let the kernel depend on parameters λ that do not enter the distribution of the random β: i.e., $L(y \mid \beta, x, \lambda)$. Denote the parameters as $\langle \theta, \lambda \rangle$, where θ is still the mean and covariance of the normally distributed coefficients. The EM recursion given in eq (1) becomes:

$$\langle \theta_{t+1}, \lambda_{t+1} \rangle = argmax_{\theta,\lambda} \sum_n \int h(\beta \mid y_n, x_n, \theta_t, \lambda_t) log[L(y \mid \beta, x, \lambda) f(\beta \mid \theta)] d\beta.$$

Unlike before, L now depends on the parameters and so does not drop out. However, L depends only on λ, and f depends only on θ, such that the two sets of parameters can be updated separately. The equivalent recursion is:

$$\theta_{t+1} = argmax_\theta \sum_n \int h(\beta \mid y_n, x_n, \theta_t, \lambda_t) log f(\beta \mid \theta) d\beta$$

as before and

$$\lambda_{t+1} = argmax_\lambda \sum_n \int h(\beta \mid y_n, x_n, \theta_t, \lambda_t) logL(y \mid \beta, x, \lambda) d\beta.$$

The latter is the MLE for the kernel model on weighted observations. If, e.g., the kernel is a logit formula, then the updated value of λ is obtained by estimating a standard (i.e., non-mixed) logit model on weighted observations, with each draw of β providing an observation. A more realistic situation is a model in which the kernel is a product of GEV probabilities (McFadden, 1978), with λ being the nesting parameters, which are the same for all agents. The updated values of the nesting parameters are obtained by MLE of the nested logit kernel on the weighted observations, where the only parameters in this estimation are the nesting parameters themselves. The parameters associated with the random coefficients are updated the same as before, as the mean and covariance of the weighted draws.

Alternative-specific constants in discrete choice models can be handled in the way just described. However, if the constants are the only parameters that enter the kernel, then the contraction suggested by Berry, Levinsohn, and Pakes (1995) can be applied rather than estimating them by ML.[5] For constants α, this contraction is a recursive application of $\alpha_{t+1} = \alpha_t + ln(\mathcal{S}) - ln(\hat{\mathcal{S}}(\theta_t, \alpha_t))$, where $\mathcal{S}$ is the sample (or population) share choosing each alternative, and $\hat{\mathcal{S}}(\theta, \alpha)$ is the predicted share based on parameters θ and α. This recursion would ideally be iterated to convergence with respect to α for each iteration of the recursion for θ. However, it is probably effective with just one updating of α for each updating of θ.

4 Application

We apply the procedure to a mixed logit model, using data on households' choice among energy suppliers in stated-preference (SP) exercises. SP exercises are often used to estimate preferences for attributes that are not exhibited in

[5]If the kernel is the logit formula, then the contraction gives the MLE of the constants, since both equate sample and predicted shares for each alternative; see, e.g., Train 2003, p. 66.

markets or for which market data provide insufficient variation for meaningful estimation. A general description of the approach, with a review of its history and applications, is provided by, e.g., Louviere et al. (2000). In an SP survey, each respondent is presented with a series of choice exercises. Each exercise consists of two or more alternatives, with attributes of each alternative described. The respondent is asked to identify the alternative that they would choose if facing the choice in the real world. The attributes are varied over situations faced by each respondent as well as over respondents, to obtain the variation that is needed for estimation.

In the current application, respondents are residential energy customers, defined as a household member who is responsible for the household's electricity bills. Each respondent was presented with 12 SP exercises representing choice among electricity suppliers. Each exercise consisted of four alternatives, with the following attributes of each alternative specified: the price charged by the supplier in cents per kWh; the length of contract that binds the customer and supplier to that price (varying from 0 for no binding to 5 years); whether the supplier is the local incumbent electricity company (as opposed to a entrant); whether, if an entrant, the supplier is a well-known company like Home Depot (as opposed to a entrant that is not otherwise known); whether time-of-use rates are applied, with the rates in each period specified; and whether seasonal rates are applied, with the rates in each period specified. Choices were obtained for 361 respondents, with nearly all respondents completing all 12 exercises. These data are described by Goett (1998). Huber and Train (2001) used the data to compare ML and Bayesian methods for estimation of conditional distributions of utility coefficients.

The behavioral model is specified as a mixed logit with repeated choices (Revelt and Train, 1998). Consumer n faces J alternatives in each of T choice situations. The utility that consumer n obtains from alternative j in choice situation t is $U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt}$, where x_{njt} is a vector of observed variables, β_n is random with distribution specified below, and ε_{njt} is iid extreme value. In each choice situation, the agent chooses the alternative with the highest utility, and this choice is observed but not the latent utilities themselves. By specifying ε_{njt} to be iid, all structure in unobserved terms is captured in the specification

of $\beta_n' x_{njt}$. McFadden and Train (2000) show that any random utility choice model can be approximated to any degree of accuracy by a mixed logit model of this form.[6]

Let y_{nt} denote consumer n's chosen alternative in choice situation t, with the vector y_n collecting the choices in all T situations. Similarly, let x_n be the collection of variables for all alternatives in all choice situation. Conditional on β, the probability of the consumer's observed choices is a product of logits:

$$L(y_n \mid \beta, x_n) = \prod_t \frac{e^{\beta' x_{ny_t t}}}{\sum_j e^{\beta' x_{njt}}}.$$

The (unconditional) probability of the consumer's sequence of choice is:

$$P(y_n \mid x_n, \theta) = \int L(y_n \mid \beta, x_n) f(\beta \mid \theta) d\beta$$

where f is the density of β, which depends on parameters θ. This f is the (unconditional) distribution of coefficients in the population. The density of β conditional on the choices that consumer n made when facing variables x_n is $h(\beta \mid y_n, x_n, \theta) = L(y_n \mid \beta, x_n) f(\beta \mid \theta) / P(y_n \mid x_n, \theta)$.

We first assume that β is normally distributed with mean b and covariance W. The recursive estimation procedure is implemented as follows, with b and W used explicitly for θ:

1. Start with trial values b_0 and W_0.

2. For each sampled consumer, take R draws of β, with the r-th draw for consumer n created as $\beta_{nr} = b_0 + C_0 \eta$ where C_0 is the lower triangular Choleski factor of W_0 and η is a vector of iid standard normal draws.

3. Calculate a weight for each draw as $w_{nr} = L(y_n \mid \beta_{nr}, x_n) / \frac{1}{R} \sum_{r'} L(y_n \mid \beta_{nr'}, x_n)$.

4. Calculate the weighted mean and covariance of the $N \cdot R$ draws, and label them b_1 and W_1.

[6] It is important to note that McFadden and Train's theorem is an existence result only and does not provide guidance on finding the appropriate distribution and specification of variables that attains a close approximation.

5. Repeat steps (2)-(4) using b_1 and W_1 in lieu of b_0 and W_0, continuing to convergence.

The last choice situation for each respondent was not used in estimation and instead was reserved as a "hold-out" choice to assess the predictive ability of the estimated models. For simulation, 200 randomized Halton draws were used for each respondent. These draws are described by, e.g., Train (2003). In the context of mixed logit models, Bhat (2001) found that 100 Halton draws provided greater accuracy than 1000 pseudo-random draws; his results have been confirmed by Train (2000), Munizaga and Alvarez-Diaziano (2001) and Hensher (2001).

The estimated parameters are given in Tables 1, with standard errors calculated as described above, using the simulated scores at convergence. Table 1 also contains the estimated parameters obtained by maximum simulated likelihood (MSLE.) The results are quite similar. Note that the recursive estimator (RE) treats the covariances of the coefficients as parameters, while the parameters for MSLE are the elements of the Choleski factor of the covariance. (The covariances are not parameters in MLE because of the difficulty of assuring that the covariance matrix at each iteration is positive definite when using gradient-based methods. By construction, the RE assures a positive definite covariance at each iteration, since each new value is the covariance of weighted draws.) To provide a more easily interpretable comparison, Table 2 gives the estimated standard deviations and correlation matrix implied by the estimated parameters for each method.

The estimated parameters were used to calculate the probability of each respondent's choice in their last choice situation. The results are given at the bottom of Table 1. Two calculation methods were utilized. First, the probability was calculated by mixing over the population density of parameters (i.e., the unconditional distribution), i.e., $P_{nT} = \int L(y_{nT} \mid \beta, x_{nT}) f(\beta \mid \hat{\theta}) d\beta$, where T denotes the last choice situation. This is the appropriate formula to use in situations for which previous choices by each sampled agent are not observed. RE gives an average probability of 0.3742, and MSLE gives 0.3620. The probability is slightly higher for RE than MSLE, which indicates that RE predicts somewhat better. The same result was observed for all the alternative

specifications discussed below. The second calculation mixes over the conditional density for each respondent, using $h(\beta \mid y, x, \hat{\theta})$ instead of $f(\beta \mid \hat{\theta})$. This formula is appropriate when previous choices of agents have been observed. The probability is of course higher under both estimators than when using the unconditional density, since each respondent's previous choices provide useful information about how they will choose in a new situation. The average probability from RE is again higher than that from MSLE. However, unlike the unconditional probability calculation, this relation is reversed for some of the alternative specifications discussed below.

The MSLE algorithm converged in 141 iterations and took 7 minutes, 4 seconds using analytic gradients and 3 hours, 20 minutes using numerical gradients.[7] For RE, I defined convergence as each parameter changing by less than one-half of one percent and the convergence statistic given above being less than 1E-4. The first of these criteria was the more stringent in this case, in that the second was met (at 0.82E-4) once the first was. RE converged in 162 iterations and took 7 minutes, 59 seconds. Since RE does not require the coding of gradients, the implication of these time comparisons is that using RE instead of MSLE reduces either the researcher's time in coding analytic gradients or the computer time in using numerical gradients.

Alternative convergence criteria were explored for RE, both more relaxed and more stringent. Using a more relaxed criterion of each parameter changing less than one percent, estimation required 63 iterations; took 3 minutes, 1 second; and obtained a convergence statistic of 1.2E-4. When the criterion was tightened to each parameter changing by less than one-tenth of one percent, estimation required 609 iterations; took 29 minutes, 3 seconds; and obtained a convergence statistic of 0.44E-4. The estimated parameters changed little by applying the stricter criterion. Interestingly, the more relaxed criterion

[7]All estimation was in Gauss on a PC with a Pentium 4 processor, 3.2GHz, with 2 GB of RAM. For MSLE, I used Gauss' maxlik routine with my codes for the mixed logit log-likelihood function and for analytic gradients under normally distributed coefficients. For RE, I wrote my own code; one of the advantages of the approach is the ease of coding it. I will shortly be developing RE routines in matlab, which I will make available for downloading from my website at http://elsa.berkeley.edu/∼train.

obtained parameters that were a bit closer to the MSL estimates. For example, the mean and standard deviation of the price coefficient were -0.927 and .611 after 62 iterations and -0.9954 and 0.5471 after 162 iterations, compared to the MSL estimates of -0.939 and 0.691.

Step-sizes are compared across the algorithms by examining the iteration log. Table 3 gives the iteration log for the mean and standard deviation of the price coefficient, which is indicative for all the parameters. The RE algorithm moves, at first, considerably more quickly toward the converged values than the gradient-based MSLE algorithm. However, it later slows down and eventually takes smaller steps than the MSLE algorithm. As Dempster et al. (1977) point out, this is a common feature of EM algorithms. However, Ruud (1991) notes that the algorithm's slowness near convergence is balanced by greater numerical stability, since it avoids the numerical problems that are often encountered in gradient-based methods, such as overstepping the maximum and getting "stuck" in areas of the likelihood function that are poorly approximated by a quadradic. We observed these problems with MSLE in two of our alternative specifications, discussed below, where new starting values were required after the MSLE algorithm failed at the original starting values. We encountered no such problems with RE.[8]

Alternative starting values were tried in each algorithm. Several different convergence points were found with each of the algorithms. All of them were similar to the estimates in Table 1, and none obtained a higher log-likelihood value. However, the fact that different converged values were obtained indicates that the likelihood function is "rippled" around the maximum. This phenomenon is not unexpected given the large number of parameters and the relatively small behavioral differences associated with different combinations of parameter values. Though this issue might constitute a warning about estimation of so many parameters, restricting the parameters doesn't necessarily

[8]The recursion can be used as an "accelerator" rather than an estimator, by using it for initial iterations and then switching to MSL near convergence. This procedure takes advantage of its larger initial steps and the avoidance of numerical problems, which usually occur in MSL further from the maximum, while retaining the familiarity of MSL and its larger step-sizes near convergence.

resolve the issue as much as mask it. In any case, the issue is the same for MSLE and RE.

Table 4 gives statistics for several alternative specifications. The columns in the table are for the following specifications:

1. All coefficients are normally distributed. This is the specification in Table 1 and is included here for comparison.

2. Price coefficient is lognormally distributed, as $-exp(\beta_p)$, with β_p and the coefficients of the other variables normally distributed. This specification assures a negative price coefficient for all agents.

3. The coefficients of price, TOU rates and seasonal rates are lognormally distributed, and the other coefficients are normal. This specification assures that all three price-related attributes have negative coefficients for all agents.

4. Price coefficient is censored normal, $min(0, \beta_p)$, with others normal. This specification prevents positive price coefficients but allows some agents to place no importance on price, at least in the range of prices considered in the choice situations.

5. Price coefficient is distributed as S_B from 0 to 2, as $2exp(\beta_p)/(1 + exp(\beta_p))$, others normal. This distribution is bounded on both sides and allows a variety of shapes within these bounds; see Train and Sonnier (2005) for an application and discussion of its use.

6. The model is specified in willingness-to-pay space, using the concepts from Sonnier et al. (2007) and Train and Weeks (2005). Utility is re-parameterized as $U = \alpha p + \alpha\beta' z + \varepsilon$ for price p and non-price attributes z, such that β is the agent's willingness to pay (wtp) for attribute z. This parameterization allows the distribution of wtp to be estimated directly. Under the usual parameterization, the distribution of wtp is estimated indirectly by estimating the distribution of the price and attribute coefficients, and deriving (or simulating) the distribution of their ratio.

MSLE and RE provide fairly similar estimates under all the specifications. In cases when the estimated mean and standard deviation of the underlying normal for the price coefficient are somewhat different, the difference is less when comparing the mean and standard deviation of the coefficient itself. For example, in specifications (2) and (3), a fifty percent difference in the estimated mean of the underlying normal translates into less than four percent difference in the mean of the coefficient itself.

For all the specifications, the log of the simulated likelihood ($\tilde{LL}$) is lower at convergence with RE than with MSLE. This difference is by construction, since the MSL estimates are those that maximize the $\tilde{LL}$, while the RE estimates are those that set the simulated scores equal to zero with the simulated scores not being the derivative of the $\tilde{LL}$. However, despite this difference, it would be useful if the $\tilde{LL}$ under the two methods moved in the same direction when the specification is changed. This is not the case. $\tilde{LL}$ is higher for specification (3) than specification (1) under either estimator. However, for specification (4), $\tilde{LL}$ under RE is higher while that under MSLE is lower than for specification (1). The change under MSLE does not necessarily provide better guidance, since simulation error can affect MSLE both in the estimates that are obtained and the calculation of the log-likelihood at those estimates.

The average probability for the "hold-out" choice using the population density is higher under RE than MSLE for all specifications. When using the conditional density, neither method obtains a higher average probability for all specifications. These results were mentioned above.

For MSLE, I used numerical gradients rather than recoding the analytic gradients. The run times in Table 4 therefore reflect equal amounts of recoding time for each method. Run times are much lower for RE than MSLE when numerical gradients are used for MSLE. With analytic gradients, MSLE would be about the same speed as RE,[9] but of course would require more coding time. As mentioned above, the ML algorithm failed for two of the specifications

[9]In some cases, MSLE is slower even with analytic gradients. For example, specification (2) was took 333 iterations in MSLE, while RE took 139. An iteration in MSLE with analytic gradients takes about the same time as an iteration in RE, such that for specification (2), MSLE with analytic gradients would be slower than RE.

(namely, 5 and 6) when using the same starting values as for the others; these runs were repeated with the converged values from specification (2) used as starting values.

5 Summary

A simple recursive estimator for random coefficients is based on the fact that the expectation of the conditional distributions of coefficients is equal to the unconditional distribution. The procedure takes draws from the unconditional distribution at trial values for its moments, weights the draws such that they are equivalent to draws from the conditional distributions, calculates the moments of the weighted draws, and then repeats the process with these calculated moments, continuing until convergence. The procedure constitutes a simulated EM algorithm and provides a method of simulated scores estimator. The estimator is asymptotically equivalent to MLE if the number of draws used in simulation rises faster than $\sqrt{N}$, which is the same condition as for MSL. In an application of mixed logit on stated-preference data, the procedure gave estimates that are similar to those by MSL, was faster than MSL with numerical gradients, and avoided the numerical problems that MSL encountered with some of the specifications.

References

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

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

Boyles, R. (1983), 'On the convergence of the em algorithm', *Journal of the Royal Statistical Society B* **45**, 47–50.

Chen, H. and S. Cosslett (1998), 'Environmental quality preference and benefit estimation in multinomial probit models: A simulation approach', *American Journal of Agricultural Economics* **80**, 512–520.

Dempster, A., N. Laird and D. Rubin (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society B* **39**, 1–38.

Goett, A. (1998), 'Estimating customer preferences for new pricing products', Electric Power Research Institute Report TR-111483, Palo Alto.

Goett, A., K. Hudson and K. Train (2000), 'Consumers' choice among retail energy suppliers: The willingnes-to-pay for service attributes', *The Energy Journal* **21**, 1–28.

Greene, W. (2000), *Econometric Analysis*, Prentice Hall, New York.

Hajivassiliou, V. and D. McFadden (1998), 'The method of simulated scores for the estimation of ldv models', *Econometrica* **66**, 863–96.

Hajivassiliou, V. and P. Ruud (1994), Classical estimation methods for ldv models using simulation, *in* R.Engle and D.McFadden, eds, 'Handbook of Econometrics', North-Holland, Amsterdam, pp. 2383–441.

Hensher, D. (2001), 'The valuation of commuter travel time savings for car drivers in new zealand: Evaluating alternative model specifications', *Transportation* **28**, 101–118.

Hensher, D., N. Shore and K. Train (2005), 'Households' willingness to pay for water service attributes', *Environmental and Resource Economics* **32**, 509–531.

Huber, J. and K. Train (2001), 'On the similarity of classical and bayesian estimates of individual mean partworths', *Marketing Letters* **12**, 259–269.

Lee, L. (1995), 'Asymptotic bias in simulated maximum likelihood estimation of discrete choice models', *Econometric Theory* **11**, 437–483.

Louviere, J. (2003), 'Random utility theory-based stated preference elicitation methods', working paper, Faculty of Business, University of Technology, Sydney.

Louviere, J., D. Hensher and J. Swait (2000), *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, New York.

McLachlan, G. and D. Peel (2000), *Finite Mixture Models*, John Wiley and Sons, New York.

Munizaga, M. and R. Alvarez-Daziano (2001), 'Mixed logit versus nested logit and probit', Working Paper, Departmento de Ingeniera Civil, Universidad de Chile.

Revelt, D. and K. Train (1998), 'Mixed logit with repeated choices', *Review of Economics and Statistics* **80**, 647–657.

Ruud, P. (1991), 'Extensions of estimation methods using the em algorithm', *Journal of Econometrics* **49**, 305–341.

Sonnier, G., A. Ainslie and T. Otter (2007), 'Hereogeneous distributions of willingness to pay in choice models', *Quantitative Marketing and Economics* **5**(3), 313–331.

Train, K. (1998), 'Recreation demand models with taste variation', *Land Economics* **74**, 230–239.

Train, K. (2000), 'Halton sequences for mixed logit', Working Paper No. E00-278, Department of Economics, University of California, Berkeley.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, New York.

Train, K. and G. Sonnier (2005), Mixed logit with bounded distributions of correlated partworths, *in* R.Scarpa and A.Alberini, eds, 'Applications of Simulation Methods in Environmental and Resource Economics', Springer, Dordrecht, pp. 117–134.

Train, K. and M. Weeks (2005), Discrte choice models in preference space and willingness-to-pay space, *in* R.Scarpa and A.Alberini, eds, 'Applications of Simulation Methods in Environmental and Resource Economics', Springer, Dordrecht, pp. 1–16.

Wu, C. (1983), 'On the convergence properties of the em algorithm', *Annals of Statistics* **11**, 95–103.

Table 1: Mixed Logit Model of Electricity Supplier Choice

All coefficients normally distributed

Recursive estimator (RE), Maximum simulated likelihood estimator (MSLE)

Parameters	RE	MSLE
	(Std errors in parentheses)	
Means		
1. Price	-0.9954 (0.0521)	-0.9393 (0.0520)
2. Contract length	-0.2404 (0.0231)	-0.2428 (0.0256)
3. Local utility	2.5464 (0.1210)	2.3328 (0.1337)
4. Well known co.	1.8845 (0.0742)	1.8354 (0.0104)
5. TOU rates	-9.3126 (0.4571)	-9.1682 (0.4400)
6. Seasonal rates	-9.6898 (0.4496)	-9.0710 (0.4365)
	Covariances	Choleski
11	0.5471 (0.0726)	0.6909 (0.0611)
21	0.0266 (0.0439)	-0.0333 (0.0290)
22	0.1222 (0.0146)	0.4180 (0.0236)
31	0.9430 (0.2672)	-1.6089 (0.1523)
32	0.3602 (0.1039)	0.2419 (0.1475)
33	2.8709 (0.3321)	1.4068 (0.1468)
41	0.5208 (0.1689)	-0.9107 (0.1218)
42	0.2668 (0.0681)	0.1526 (0.1192)
43	1.3065 (0.3543)	0.6746 (0.1216)
44	1.1015 (0.1339)	-1.0424 (0.0997)
51	4.5204 (1.2492)	-4.6228 (0.4740)
52	0.2707 (0.3972)	-0.1813 (0.1690)
53	7.9995 (2.4263)	1.8399 (0.1740)
54	4.5092 (1.5356)	0.3592 (0.2026)
55	45.050 (5.9201)	2.6309 (0.1631)
61	4.4860 (1.1875)	-5.3688 (0.4862)
62	0.1156 (0.3933)	-0.3913 (0.1302)
63	7.5672 (2.2439)	0.4850 (0.1691)
64	3.8878 (1.3968)	0.5309 (0.2054)
65	39.927 (10.578)	1.1074 (0.1544)
66	41.916 (5.2169)	1.7984 (0.1371)
Log of Sim. Likelihood	-3482.93	-3423.08
Average probability of chosen alt. in last situation.		
Unconditional density	0.3742	0.3620
Conditional density	0.5678	0.5632

Table 2: Standard deviations and correlations

	Std devs		Correlations					
	RE	MSLE	RE bottom, MSLE top					
Price	0.740	0.691	1.000	0.079	0.748	0.589	0.819	0.921
Contract	0.350	0.419	0.103	1.000	0.172	0.145	0.033	0.006
Local util	1.694	2.151	0.752	0.608	1.000	0.736	0.822	0.736
Well known	1.050	1.547	0.671	0.728	0.735	1.000	0.578	0.511
TOU rates	6.712	5.643	0.911	0.115	0.703	0.640	1.000	0.879
Seasonal	6.474	5.827	0.937	0.051	0.690	0.572	0.919	1.000

Table 3: Iterations

Price coefficents

Iteration	Mean		Std dev	
	RE	MSLE	RE	MSLE
1	0	0	2.449	0.2000
2	0.1431	0.1108	0.6650	0.1259
3	0.0479	0.0718	0.3641	0.1567
4	-0.0553	0.0174	0.2944	0.1120
5	-0.1405	0.1127	0.2657	0.2430
6	-0.2136	0.0567	0.2567	0.2217
7	-.2762	-0.0326	0.2575	0.1552
8	-.3293	-0.0162	0.2625	0.1717
9	-.3746	-0.0070	0.2702	0.1687
10	-.4132	-0.0064	0.2800	0.1514
20	-.6416	-0.3322	0.4191	0.0697
30	-0.7607	-0.5693	0.5346	0.3825
40	-0.8357	-0.7316	0.5947	0.4505
50	-0.8869	-0.7919	0.6325	0.5633
60	-0.9217	-0.8913	0.6573	0.6173
70	-0.9446	-0.9272	0.6738	0.6414
80	-0.9602	-0.9399	0.6854	0.6559
90	-0.9711	-0.9325	0.6941	0.6485
100	-0.9776	-0.9415	0.7006	0.6593
110	-0.9827	-0.9462	0.7044	0.6688
120	-0.9856	-0.9464	0.7087	0.6786
130	-0.9848	-0.9425	0.7178	0.6862
140	-0.9832	-0.9396	0.7282	0.6904
150	-0.9862	NA	0.7341	NA
160	-0.9932	NA	0.7381	NA

Table 4: Alternative Specifications

	All normal	Price log normal	Price TOU season lognorm	Price censor normal	Price S_B	WTP space
Price						
Underlying normal						
Mean						
RE	-.9954	-.2441	-.1692	-1.0203	-.1761	-0.0892
MSLE	-.9393	-.1655	-.2466	-.9828	-.0915	-.1125
Std dev						
RE	.7397	.5560	.6274	.6253	1.316	.2941
MSLE	.6909	.4475	.7903	.6530	1.7964	.2444
Coefficient						
Mean						
RE	-.9954	-.9144	-1.028	-1.033	-.9335	-.9551
MSLE	-.9393	-.9397	-1.068	-1.002	-.9711	-.9207
Std dev						
RE	.7397	.5503	.7140	.5971	.4990	.2871
MSLE	.6909	.4411	.9946	.6155	.5958	.2284
Probability						
for last choice						
Population density						
RE	.3742	.3629	.3702	.3785	.3688	.3696
MSLE	.3620	.3557	.3539	.3565	.3662	.3649
Conditional densities						
RE	.5678	.5501	.5640	.5630	.5634	.5309
MSLE	.5632	.5569	.5674	.5691	.5637	.5415
Log Sim. Likelihd						
RE	-3482.93	-3510.81	-3467.49	-3508.84	-3474.66	-3554.66
MSLE	-3423.08	-3456.63	-3420.58	-3420.21	-3424.19	-3494.48
Run time						
RE	7m59s	6m45s	12m2s	11m27s	10m14s	22m32s
MSLE*	3h20m30s	7h54m34s	3h31m31s	6h26m46s	2h51m9s	4h6m5s

*Using numerical derivatives. Staring values for (5) and (6) are estimates from (2).