JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **15**: 447–470 (2000)

MIXED MNL MODELS FOR DISCRETE RESPONSE

DANIEL McFADDEN* AND KENNETH TRAIN
Department of Economics, University of California, Berkeley, CA, 94720-3880, USA

SUMMARY

This paper considers mixed, or random coefficients, multinomial logit (MMNL) models for discrete response, and establishes the following results. Under mild regularity conditions, any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a MMNL model. Practical estimation of a parametric mixing family can be carried out by Maximum Simulated Likelihood Estimation or Method of Simulated Moments, and easily computed instruments are provided that make the latter procedure fairly efficient. The adequacy of a mixing specification can be tested simply as an omitted variable test with appropriately defined artificial variables. An application to a problem of demand for alternative vehicles shows that MMNL provides a flexible and computationally practical approach to discrete response analysis. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Define a *mixed multinomial logit (MMNL)* model as a MNL model with *random coefficients* α drawn from a cumulative distribution function $G(\alpha;\theta)$:

$$P_{\mathrm{C}}(i \mid \boldsymbol{x}, \theta) = \int L_{\mathrm{C}}(i; \boldsymbol{x}, \alpha) \cdot G(\mathrm{d}\alpha; \theta) \text{ with } L_{\mathrm{C}}(i; \boldsymbol{x}, \alpha) = e^{x_i \alpha} / \sum_{j \in C} e^{x_j \alpha} \tag{1}$$

In this setup, $\mathbf{C} = \{1, \ldots, J\}$ is the choice set; the x_i are $1 \times K$ vectors of functions of observed attributes of alternative i and observed characteristics of the decision maker, with $\boldsymbol{x} = (x_1, \ldots, x_J)$; α is a $K \times 1$ vector of random parameters; $L_{\mathrm{C}}(\mathrm{i}; \boldsymbol{x}, \alpha)$ is a MNL model for the choice set $\mathbf{C}$; and θ is a vector of deep parameters of the mixing distribution G. The random parameters α may be interpreted as arising from taste heterogeneity in a population of MNL decision makers. If the x_i contain alternative-specific variables, then the corresponding components of α can be treated as alternative-specific random effects. Alternately, the model may simply be interpreted as a flexible approximation to choice probabilities generated by a random utility model. The mixing distribution G may come from a continuous parametric family, such as multivariate normal or log normal, or it may have a finite support. When G has finite support, MMNL models are also called *latent class* models. Equation (1) describes a single decision, but extension to dynamic choice models with multiple decisions is straightforward, by mixing over the parameters of a product of MNL models for each component decision.

* Correspondence to: Daniel McFadden, Department of Economics, University of California, Berkeley, CA 94720-3880, USA; e-mail: mcfadden@econ.berkeley.edu

Contract/grant sponsor: E. Morris Cox Fund.

Received 30 November 1998
Revised 15 May 2000

The MMNL model was introduced by Boyd and Mellman (1980) and Cardell and Dunbar (1980), although an earlier literature had considered the mathematically similar problem of aggregating the MNL model over a distribution of explanatory variables; see Talvitie (1972), Westin (1974), McFadden and Reid (1975), and Westin and Gillen (1978). There is a lengthy literature investigating various aspects of the MMNL model; see Beggs (1988), Börsch-Supan (1990), Brownstone and Train (1999), Chavas and Segerson (1986), Dubin and Zeng (1991), Enberg, Gottschalk, and Wolf (1990), Follman and Lambert (1989), Formann (1992), Gonul and Srinivasan (1993), Jain, Vilcassim, and Chintagunta (1994), Montgomery, Richards, and Braun (1986), Reader (1993), Revelt and Train (1998), Steckel and Vanhonacker (1988), Train, McFadden, and Goett (1987), and Train (1998, 1999). Chesher and Santos-Silva (1995) have developed specification tests for MMNL that are relatives of ones proposed here. This paper establishes the following results:

- Under mild regularity conditions, MMNL models are random utility maximization (RUM) models, and any discrete choice model derived from a RUM model has choice probabilities that can be approximated as closely as one pleases by a MMNL model (Section 2).
- Numerical integration or approximation by simulation is usually required to evaluate MMNL probabilities. *Maximum Simulated Likelihood* (MSLE) or *Method of Simulated Moments* (MSM) can be used to estimate the MMNL model (Section 3).
- The adequacy of a mixing specification can be tested simply as an omitted variable test with appropriately defined artificial variables (Section 4).
- An application to a problem of demand for alternative vehicles shows that MMNL provides a flexible and computationally practical approach to discrete response analysis (Section 5).

2. A GENERAL APPROXIMATION PROPERTY OF MMNL

Economic theory often suggests that discrete responses are the result of optimization of payoffs to decision makers: utility for consumers, profit for firms. The following discussion will be phrased in terms of utility-maximizing consumers. When unobserved heterogeneity in the population of consumers is accounted for, this will lead to a class of response models based on *random utility maximization* (RUM). A resource allocation to a consumer will specify quantities of goods and leisure, and for our particular interest the attributes of a discrete alternative, such as an automobile model. We will consider two sources of unobserved heterogeneity: features of alternatives that are not recorded by the analyst, and unmeasured consumer characteristics that determine preferences.

Let $q = (g, l, z, \zeta)$ denote a consumer's resource allocation, where z is a vector of observed attributes and ζ is a vector of unobserved attributes of a discrete alternative, g is a vector of quantities of other goods, and l is leisure. Assume that the domain of q is a compact rectangle in a finite-dimensional Euclidean space. Consumers have a vector of observed characteristics s and a vector of unobserved characteristics ς; with (s,ς) determining preferences over resource allocations. Assume that the domain of (s,ς) is a compact subset of a finite-dimensional Euclidean space. This is not a substantive restriction for discrete choice analysis when the number of choice alternatives is bounded. Assume that consumer preferences over resource allocations, $\succeq_{s,\varsigma}$, are complete and transitive, with the continuity property that if a sequence of allocations and consumer characteristics converges, $(q^{1\prime}, q^{2\prime}, s', \varsigma') \to (q^1, q^2, s, \varsigma)$, and satisfies $q^{1\prime} \succeq_{s',\varsigma'} q^{2\prime}$, then $q^1 \succeq_{s,\varsigma} q^2$. For fixed (s,ς), this is the standard continuity condition on preferences. Our

condition extends this to require that consumers with similar characteristics will also have similar preferences. Together, these assumptions imply that preferences can be represented by a utility function $U(g, l, z, \zeta, s, \varsigma)$ that is continuous in its arguments; see Appendix Lemma 1.

We next consider the stochastic properties of unobserved elements in this formulation of the consumer's problem. Let $(\Omega,\mathcal{W},\pi)$ denote a fundamental probability space, where $\mathcal{W}$ is the σ-field of measurable subsets and π is a probability measure. Let $\mathbf{T}$ denote a subset of $\mathbb{R}^m$, and $X : \Omega \times \mathbf{T} \to \mathbb{R}^n$ denote a *continuous random field*; i.e., for each $t \in \mathbf{T}$, $X(\cdot, t)$ is a random vector, measurable with respect to $\mathcal{W}$, and for a set of ω occurring with probability one, $X(\omega, \cdot)$ is a continuous function on $\mathbf{T}$. We will often suppress the dependence of the random field on ω, and write it as $X(t)$. A continuous random field has $\lim_{t' \to t} X(t') = X(t)$ with probability one, implying that $X(t')$ converges in distribution to $X(t)$ as $t' \to t$. The CDF of $X(t)$ is $F(x, t) = \pi(\{\omega \in \Omega \mid X(\omega, t) \leqslant x\})$. We say that X has a *regular canonical representation* if there exists a continuous function $h : [0, 1]^n \times \mathbf{T} \to \mathbb{R}^n$ and a uniformly distributed continuous random field $\varepsilon : \Omega \times \mathbf{T} \to [0, 1]^n$ such that $X(t) = h(\varepsilon(t), t)$ with probability one.[1] We show in the Appendix that a continuous random field whose CDF admits a positive continuous density has a regular canonical representation. For example, if $X(t)$ is a mean-zero Gaussian continuous random field with a definite covariance matrix $\Omega(t)$, and Φ denotes the standard normal CDF, then $\Omega(t)$ has a continuous Cholesky factor $\Lambda(t)$ and the mapping $\varepsilon(t) = \Phi(\Lambda(t)^{-1}X(t))$ is a uniformly distributed continuous random field that inverts to the regular canonical representation $X(t) = \Lambda(t)\Phi^{-1}(\varepsilon(t))$.

A primitive postulate of preference theory is that tastes are established prior to assignment of resource allocations. Then, the distribution of ς cannot depend on q, although in general it will depend on s. We assume that $\varsigma = \varsigma(s)$ is a continuous random field with a regular canonical representation, and write it as $\varsigma(s) = h_0(\upsilon(s), s)$, where $\upsilon(s)$ is a uniformly distributed continuous random field. Then consumers with similar observed characteristics will have similar distributions of unobserved characteristics. Another primitive postulate of consumer theory is that the description of a resource allocation does not depend on consumer characteristics. Thus, consumers' tastes and perceptions do not enter the 'objective' description of a resource allocation, although they will obviously enter the consumer's evaluation of the allocation. This postulate implies that the distribution of ξ cannot depend on (s,υ), although it may depend on z. We will assume that ζ is specified as a continuous random field with a regular canonical representation, and write it as $\zeta(z) = h(\varepsilon(z), z)$, where $\varepsilon(s)$ is a uniformly distributed continuous random field. Then discrete alternatives that are similar in their observed attributes will have similar distributions of unobserved attributes. Substituting the transformations h_0 and h into the definition of U, we can consider a *canonical random utility model* $U(g, l, z, s, \varepsilon(z), \upsilon(s))$ that is continuous in its arguments, with $\varepsilon(z)$ and $\upsilon(s)$ independently uniformly distributed continuous random fields.

Economic consumers make choices subject to dollar and time budgets. For discrete choice, if assigned a discrete alternative z, the consumer will choose goods g and leisure l to maximize utility subject to these budgets. If the alternative requires time $t = \tau(z)$, the consumer's 24-hour/day time budget is $24 = l^* + e + t$, where e is hours worked and l^* is hours of pure leisure. If only a portion λ of the time t devoted to the alternative is equivalent to work, then $l = l^* + (1 - \lambda)t$ is the effective leisure entering the utility function. Suppose the consumer faces a dollar budget

$$a + w \cdot e = p \cdot g + c \tag{2}$$

[1] A random field ε is *uniformly distributed* if $\varepsilon(t)$ has a uniform distribution on $[0,1]^n$ for each $t \in \mathbf{T}$; see Appendix Lemma 3.

where a is non-wage income, c is the cost of the discrete alternative, w is the wage, and p is the vector of goods prices. For the assigned alternative, maximum utility then satisfies

$$U'(a-c, p, w, t\lambda, z, s, \varepsilon(z), \upsilon(s)) = \max_{e,g} U(g, 24-e-t\lambda, z, s, \varepsilon(z), \upsilon(s))$$
$$\text{subject to } a + w{\cdot}e = p{\cdot}g + c$$

This is a *conditional indirect utility function*, given the discrete alternative.[2] With a monotone transformation, we can assume that the range of utility is contained in the unit interval. For economic applications, it will be important to distinguish the market variables a, p, and w, which can be altered by economic policy, from the observed consumer characteristics included in s. Similarly, the market cost c of the discrete alternative which can be altered by policy is distinguished from z, while t is a component of z. An important implication of these distinctions is that for each realization of $\varepsilon(z)$ and $\upsilon(s)$, the conditional indirect utility function is characterized by the standard economic properties that it is increasing in $a-c$, non-increasing in (p,w), and homogeneous of degree zero and quasi-convex in $(a-c, p, w)$. It will be convenient as a shorthand in the following analysis to redefine z and s to absorb the market variables, and write the conditional indirect utility function as $U(z, s, \varepsilon(z), \upsilon(s))$, keeping in mind that $\varepsilon(z)$ and υ(s) will not depend on the market variable components of z and s. Let $\mathbf{Z}$ and $\mathbf{S}$ denote the domains of z and s, respectively, and note that they are assumed to be compact subsets of finite-dimensional spaces.

Consider choice over finite sets of discrete alternatives $\mathbf{C} = \{z_1, \ldots, z_J\}$, distinguished by the consumer (and the observer) in terms of their observed attributes z_j which may include 'brand names' or other alternative-specific identifiers that influence the consumer's evaluation. We will interpret $\mathbf{C}$ as an ordered sequence, and denote the family of possible J-element choice sets by $\mathcal{C}_J \subset \mathbf{Z}^J$. By construction, all the elements of $\mathbf{C} \in \mathcal{C}_J$ must be distinct. We assume that $\mathcal{C}_J$ is compact; this excludes cases where alternatives are observationally indistinguishable in the limit. We assume that there is an upper bound $\mathbf{J}^*$ on the number of elements in a choice set. Then $\mathcal{C}^* = \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_{J^*}$ is the universe of possible choice sets. We assume that proper subsets of possible choice sets are also possible; in particular, if $\mathbf{C} \in \mathcal{C}_J$ contains elements z' and z'', then $\{z', z''\} \in \mathcal{C}_2$. For brevity, $\mathbf{C} = \{z_1, \ldots, z_J\}$ will sometimes be written as $\mathbf{C} = \{1, \ldots, J\}$.

In a well-specified RUM model, there will be zero probability of ties in a choice set $\mathbf{C} = \{z_1, \ldots, z_J\}$, so that a realization $\upsilon = \upsilon(s)$ and $\varepsilon_j = \varepsilon(z_j)$ for $j = 1, \ldots, J$ of the random elements in the model almost surely determines a unique choice. When U is continuously differentiable, a sufficient condition for this is that the Jacobian

$$\begin{bmatrix} \partial U(z_1, s, \varepsilon_1, \upsilon)/\partial \upsilon & \partial U(z_1, s, \varepsilon_1, \upsilon)/\partial \varepsilon_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \partial U(z_J, s, \varepsilon_J, \upsilon)/\partial \upsilon & 0 & \cdots & \partial U(z_J, s, \varepsilon_J, \upsilon)/\partial \varepsilon_J \end{bmatrix}$$

have rank at least $J-1$, and that the support of $(\upsilon, \varepsilon_1, \ldots, \varepsilon_J)$ contain the space spanned by the Jacobian. Ways to guarantee no ties include taste factors (determined by υ) of the required dimension that interact with a full-rank array of alternative attributes, or a full set of alternative-specific effects (determined by the ε_j), or some combination. The following result establishes that

[2] The conditional indirect utility function (3) is modified in obvious ways if the consumer cannot choose work hours e, the time requirement t for the discrete alternative is absent, or time required for consumption of other goods enters the time budget.

MNL mixtures can closely approximate a very broad class of RUM models that have zero probability of ties:

Theorem 1. Let $z \in \mathbf{Z}$, with $\mathbf{Z}$ compact, denote the vector of observed attributes of a discrete alternative, and $s \in \mathbf{S}$, with $\mathbf{S}$ compact, denote the vector of observed characteristics of the consumer. Suppose discrete choices are made from choice sets $\mathbf{C} = \{z_1, \ldots, z_J\}$, with at most J^* alternatives, contained in a compact universe $\mathcal{C}^*$ in which all alternatives are distinct. Let $\mathbf{z} = (z_1, \ldots, z_J)$, and as a shorthand let $\mathbf{C} = \{1, \ldots, J\}$. Suppose discrete responses maximize a canonical conditional indirect utility function $U^*(z_j, s, \varepsilon_j, \upsilon)$ that is a bounded continuous function of its arguments, where $\varepsilon_j = \varepsilon(z_j)$ and $\upsilon = \upsilon(s)$ are uniformly distributed continuous random fields. Assume there is zero probability of ties. Let $P^*_{\mathbf{C}}(i \mid \mathbf{z}, s)$ denote the choice probabilities generated by maximization of U^* over $\mathbf{C}$. If η is a small positive scalar, then there exists a continuous function $x = x(z, s)$ of dimension $1 \times k$ for some integer k, with $\mathbf{x} = (x(z_1, s), \ldots, x(z_J, s))$, and a random utility model with choice probabilities $P_{\mathbf{C}}(i \mid \mathbf{x}, \theta)$ of the MMNL form (1), such that $P^*_{\mathbf{C}}(i \mid \mathbf{z}, s)$ and $P_{\mathbf{C}}(i \mid \mathbf{x}, \theta)$ differ by at most η for all $s \in \mathbf{S}$ and $\mathbf{z} \in \mathcal{C}^*$.

The proof is given in the Appendix. The construction in the proof shows that the random coefficients α in equation (1) can be taken to be continuous polynomial transformations of the uniformly distributed continuous random fields $\varepsilon(z)$ and $\upsilon(s)$, and from the earlier discussion the indexing of these fields will exclude economic market variables. Then, the distribution of α will not depend on observed variables except through the correlations across similar alternatives. One implication of the theorem is that MMNL can be used to approximate computationally difficult parametric random utility models simply by taking the distributions underlying these models, suitably scaled, as the mixing distributions. These can be interpreted as simulation approximations using a MNL kernel. For multinomial probit models, Brownstone and Train (1999) and Ben-Akiva and Bolduc (1996) find in Monte Carlo experiments that MMNL gives approximations that are as accurate and quick as direct simulation alternatives such as the Geweke–Hajivassiliou–Keane (GHK) simulator; see Hajivassiliou and Ruud (1994).

The theorem is stated for a single choice, but applies by reinterpretation to multiple or dynamic choice applications by treating each possible portfolio of choices as a distinct alternative. Alternately, the theorem extends easily to a time series of serially correlated RUM models approximated by serially correlated mixtures of a product of MNL models for the individual decisions: MMNL probabilities for a sequence of choices i_t from sets $\mathbf{C}_t$ for $t = 1, \ldots, T$ will have the form $P_{\mathbf{C}}(i_1, \ldots, i_T \mid \mathbf{x}, \theta) = \int \prod_{t \leqslant T} L_{Ct}(i_t; \mathbf{x}_t, \alpha) \cdot G(\mathrm{d}\alpha; \theta)$, where $L_{\mathbf{C}t}(i_t; \mathbf{x}_t, \alpha)$ is a MNL probability for period t choice, with $\mathbf{x}_t$ a vector of functions of alternative attributes and consumer characteristics that may include state dependence on historical choices, and G is a distribution that can in general include the effects of unobserved heterogeneity and serial correlation. When both state dependence and unobserved heterogeneity are present, this model suffers from Heckman's initial values problem, and a latent class form of the model with G depending on initial state can be interpreted as the Heckman–Singer semiparametric treatment for this problem; see Heckman (1981), Heckman and Singer (1984, 1986), and Heckman, Lochner, and Taber (1998). In particular, a MMNL model in this form can approximate a dynamic choice model generated by a RUM model with a multivariate normal distribution of unobserved factors.

In the proof of the theorem, a polynomial approximation to the true random utility function is perturbed by adding scaled i.i.d. Extreme Value Type I disturbances ν, yielding MNL as the base model to which mixing is applied. At this step, one could have used other distributions for the ν, although most alternatives are not as computationally tractable as MNL. For example, one might take the ν to be scaled i.i.d. standard normal. When the mixing distribution is multivariate normal, this can be interpreted as the method for simulation of the MNP model proposed by Stern (1994). Adopting i.i.d. standard normals for the base model adds one dimension of numerical integration, and requires computation of a product of univariate normal CDF's for each integration point and each decision maker. This requires more computation than a MNL base model; see Train (1995). One can use classical orthogonal polynomials, Fourier series, neural nets, or wavelets as a *basis* $x_k(z, s), k = 1, 2, \ldots$ for the approximation. Judicious choice of a basis can make the approximation more parsimonious and easier to identify econometrically than simple polynomials, and may make it easier to impose or check monotonicity and quasi-convexity properties of a conditional indirect RUM. In applications, it is often desirable to make the leading terms in the basis expressions that occur in standard parametric economic consumer models such as a Stone–Geary specification. Then, a satisfactory approximation may be achieved without a large number of additional terms.

There are two approximation results available in the literature that are somewhat different from Theorem 1. Discrete choice models continuous in their arguments can be approximated by MNL models in which the scale value of each alternative is a general function of all variables for all choices; see McFadden (1984). This approximation, sometimes called 'mother logit', does not require that the discrete choice model come from a random utility model, and the MNL approximation is not guaranteed to be consistent with RUM. Thus, this approximation can be useful for testing a RUM/MNL specification against alternative models that are not necessarily RUM, but is not useful for approximations within the RUM family. Dagsvik (1994) establishes, for a general class of RUM that have a representation in which the random effect is additive and independent of (z, s), that the random utility process can be approximated by a generalized extreme value process. Specialized to the current problem, this shows that this class of random utility models can be approximated by generalized Extreme Value RUM. This is a powerful theoretical result, but its practical econometric application is limited by the difficulty of specifying, estimating, and testing the consistency of relatively abstract generalized Extreme Value RUM.

One limitation of Theorem 1 is that it provides no practical indication of how to choose parsimonious mixing families, or how many terms are needed to obtain acceptable approximations to $P_C(i \mid \mathbf{z}, s)$. However, Monte Carlo studies indicate that fairly simple mixing structures, with random coefficients following a factor analytic structure of relatively low dimension, and relatively simple mixing families, such as latent class models with relatively few classes, are sufficiently flexible to capture quite complex patterns of heterogeneity; see Bolduc, Fortin and Gordon (1996) and Brownstone and Train (1999). The specification tests described in Section 4 are one practical adaptive approach to obtaining satisfactory approximations. In principle, one can combine a method of sieves for specification of the x_i variables with a latent class structure for the mixing distribution G to develop a fully non-parametric approach to estimation of random utility models for discrete choice.

A second limitation of the theorem is that while it guarantees the existence of a satisfactory MMNL approximation, it leaves open the possibility that identification conditions for regular maximum likelihood estimates of the MMNL model may fail, or that estimates may blow up. The first possibility is the usual local and global identification problem, reduced but not eliminated by judicious choice of the basis and careful global search in estimation. The second possibility of estimates blowing up arises if the linear approximation and mixing distribution happen to be *exact*, so that the true random utility model satisfies $U^*(z_i, s, \varepsilon, \upsilon) \equiv x_i \cdot \alpha$ with x_i a vector of polynomials in z_i and s and α distributed $G(\alpha; \theta)$. Then by scaling down the i.i.d. Extreme Value perturbations to $U^*(z_i, s, \varepsilon, \upsilon)$, one can make the MMNL approximation converge to $P^*_C(i \mid \mathbf{z}, s)$. This corresponds to approaching the maximum likelihood by scaling the MNL coefficients by a factor $\psi \to \infty$ in $P_C(i \mid \mathbf{x}, \theta, \psi) = \int L_C(i \mid \mathbf{x}, \alpha \cdot \psi) \cdot G(\mathrm{d}\alpha; \theta)$; a finite maximand does not exist. This is rarely a practical problem, since any specification of x and G adopted in an application will almost certainly miss features of the true random utility model, and ψ will be determined by a search to achieve a best approximation to the influence of these omitted factors. Alternatively, if the exact model contains additive i.i.d. Extreme Value I components, the problem cannot arise. Suppose a random utility model $U^*(z, s, \varepsilon, \upsilon)$ and $\mathbf{C} = \{z_1, \ldots, z_J\}$. Let $F(u \mid \mathbf{z}, s)$ be the CDF of $(U^*(z_1, s, \cdot), \ldots, U^*(z_J, s, \cdot))$. A necessary and sufficient condition for additive i.i.d. Extreme Value I components is that $F(-\log(t_1), \ldots, -\log(t_J))$ have the properties of a multivariate Laplace Transform: derivatives of all orders with

$$(-1)^{n_1+\cdots+n_J} \cdot \partial^{n_1+\cdots+n_J} F(-\log(t_1), \ldots, -\log(t_J)) / \partial^{n_1} t_1 \ldots \partial^{n_J} t_J \geqslant 0$$

See Appendix Lemma 4. In practice, it is difficult to find CDFs satisfying this condition, and difficult to test the condition, so that the likely possibility that the model is not exact is the best guarantee for convergence of estimators.

3. SIMULATION OF THE MMNL MODEL

A tractable empirical form for the MMNL model $P_C(i \mid \mathbf{x}, \theta) = \int L_C(i; \mathbf{x}, \alpha) \cdot G(\mathrm{d}\alpha; \theta)$ is obtained by taking $\alpha = \beta + \Lambda\zeta$, where β is a $K \times 1$ vector of 'mean' coefficients, Λ is a $K \times M$ matrix of factor loadings, with exclusion restrictions for identification, and ζ is an $M \times 1$ vector of factor levels that are independently distributed with a 'standard' density $f(\zeta)$. (This specification includes models with alternative-specific random effects: take x_j to include alternative-specific dummies and introduce factors that load on these dummies.) Let vec(B) denote the operation that stacks the columns of an array B into a vector, define $\gamma = \mathrm{vec}(\Lambda')$, let $\theta' = (\beta', \gamma')$ denote the vector of parameters of this model, and let θ_0 denote the true value of θ. Define $x_C(\zeta) = \Sigma_{j \in C} x_j \cdot L_C(j; \mathbf{x}, \beta + \Lambda\zeta)$ and let $x_{iC}(\zeta) = x_i - x_C(\zeta)$. Let $\mathbf{E}_{\zeta \mid i}$ denote an expectation with respect to the density of ζ conditioned on the event that i is chosen; i.e., the density $L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \cdot f(\zeta) / \int L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \cdot f(\zeta) \mathrm{d}\zeta$. Then $P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_\zeta L_C(i; \mathbf{x}, \beta + \Lambda\zeta)$, $\nabla_\beta \log P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_{\zeta \mid i} x_{iC}(\zeta)'$ and $\nabla_\gamma \log P_C(i \mid \mathbf{x}, \theta) = \mathrm{vec}(\mathbf{E}_{\zeta \mid i} \zeta x_{iC}(\zeta))$.

3.1. Simulation of the MMNL Probabilities and Their Derivatives

If the integral $P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_\zeta L_C(i; \mathbf{x}, \beta + \Lambda\zeta)$ can be obtained analytically, or by computationally feasible numerical integration of low dimension, then conventional maximum likelihood can be used to estimate θ. Otherwise, it is possible to simulate $P_C(i \mid \mathbf{x}, \theta)$ and its derivatives, and use

these simulation approximations for statistical inference. Make Monte Carlo draws $\zeta_p,\ p = 1, \ldots, r$, from $f(\zeta)$. Let $\mathbf{E}_r$ denote an empirical expectation with respect to a simulation sample of size r. Then,

$$P_C^r(i \mid \mathbf{x}, \theta) = \sum_{p=1}^{r} L_C(i; \mathbf{x}, \beta + \Lambda\zeta_p) \equiv \mathbf{E}_r L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \tag{4}$$

is a positive, unbiased estimator of $P_C(i \mid \mathbf{x}, \theta)$ that is continuous and continuously differentiable to all orders in θ. The derivatives of log $P_C(i \mid \mathbf{x}, \theta)$ involve conditional expectations $\mathbf{E}_{\zeta \mid i} b(\zeta) \equiv \{\mathbf{E}_\zeta b(\zeta) \cdot L_C(i; \mathbf{x}, \beta + \Lambda\zeta)\}/P_C(i \mid \mathbf{x}, \theta)$ for various functions $b(\zeta)$. These expectations are simulated by

$$E_{r \mid i} b(\zeta) \equiv \{\mathbf{E}_r b(\zeta) \cdot L_C(i; \mathbf{x}, \beta + \Lambda\zeta)\}/P_C^r(i \mid \mathbf{x}, \theta) \tag{5}$$

which is again continuously differentiable to all orders in θ. This can be interpreted as importance sampling with draws from $f(\zeta)$ as the comparison density. The simulator $\mathbf{E}_{r \mid i} b(\zeta)$ is not unbiased in general because of the appearance of the simulator $P_C^r(i \mid \mathbf{x}, \theta)$ in the denominator. Similarly, the simulator log $P_C^r(i \mid \mathbf{x}, \theta)$ of log $P_C(i \mid \mathbf{x}, \theta)$ is not unbiased because of the non-linear transformation. However, all the simulators above are consistent when $r \to \infty$. It is possible to get unbiased, but no longer continuous, estimates of $E_{r \mid i} b(\zeta)$ using an acceptance/rejection procedure that accepts draws that would produce i as the choice. Some computations require the second derivatives of log $P_C(i \mid \mathbf{x}, \theta)$; these are given in Appendix Lemma 5. In applications, these second derivatives can alternately be obtained by numerical differentiation of the formulas for the first derivatives.

In the statistical procedures to be discussed next, it will be critical that the simulators satisfy a condition of *stochastic equicontinuity*, which requires that they not 'chatter' as θ changes. This is easily accomplished by keeping the draws ζ_p fixed during iterative procedures that adjust θ; this can be done by storing the ζ_p or by regenerating them from fixed seeds.

3.2. Maximum Simulated Likelihood Estimation (MSLE)

Maximum Simulated Likelihood Estimation (MSLE) finds an estimator θ_N that maximizes the simulated log likelihood, $\mathbf{E}_N \log P_C^r(i \mid \mathbf{x}, \theta)$, with $\mathbf{E}_N$ denoting empirical expectation for a random sample of size N. Hajivassiliou and McFadden (1997) show that under mild regularity conditions, a stochastic equicontinuity property, and $r \cdot N^{-1/2} \to \infty$ as $N \to \infty$, the MSLE estimator θ_N is asymptotically equivalent to the classical maximum likelihood estimator. However, estimators that are relatively free of simulation bias in moderate samples are likely to require r considerably larger than $N^{1/2}$. Monte Carlo draws need not be independent across observations, or across the simulators of different derivatives that may be used in iterative search for θ_N. It is also possible to allow dependence across the different simulation draws, provided there is sufficient mixing for them to satisfy a central limit property. In particular, Train (1999) has found that patterned pseudo-random numbers such as Halton sequences give estimators that in Monte Carlo studies give lower mean square errors than independent random draws. We give an estimator for the asymptotic covariance matrix of θ_N only for the case of independent simulators across observations. Define the arrays

$$\Gamma_N(\theta) = -\mathbf{E}_N \nabla_{\theta\theta'} \log P_C^r(i \mid \mathbf{x}, \theta) \text{ and } \Delta_N(\theta) = \mathbf{E}_N\{\nabla_\theta \log P_C^r(i \mid \mathbf{x}, \theta)\}\{\nabla_\theta \log P_C^r(i \mid \mathbf{x}, \theta)\}' \quad (6)$$

As $r \to \infty$, both $\Gamma_N(\theta_N)$ and $\Delta_N(\theta_N)$ converge to $\Omega_N(\theta_o) = \mathbf{E}_N\{\nabla_\theta \log P_C(i \mid \mathbf{x}, \theta_o)\}\{\nabla_\theta \log P_C(i \mid \mathbf{x}, \theta_o)\}'$, so that $\Gamma_N(\theta_N)^{-1}$ and $\Delta_N(\theta_N)^{-1}$ are consistent estimators of the asymptotic covariance estimator. However, for finite r, $\Delta_N(\theta_N)$ is larger than $\Gamma_N(\theta_N)$ due to simulation noise, and $\Delta_N(\theta_N)$ decreases as r increases. Consequently, $\Delta_N(\theta_N)^{-1}$ may substantially underestimate the covariance of the estimator when r is finite, and may suggest erroneously that increasing r decreases the precision of the estimator. For this reason, we recommend the *robust asymptotic covariance matrix estimator* $\Gamma_N(\theta_N)^{-1}\Delta_N(\theta_N)\Gamma_N(\theta_N)^{-1}$ that is associated with quasi-maximum likelihood estimation; see Newey and McFadden (1994, p. 2160).

3.3. Method of Simulated Moments

Let d_i denote an indicator that is one when i is chosen, zero otherwise, and let $\mathbf{d} = (d_1, \ldots, d_J)$. A classical method of moments estimator for θ can be based on the condition that the *generalized residual* $d_i - \mathbf{E}_\zeta L_C(i; \mathbf{x}, \beta + \Lambda\zeta)$, evaluated at the true parameters, is orthogonal in the population to any *instrument vector* $W_i(\mathbf{x}, \theta)$ that has the dimension of θ. Write this moment as

$$\mathrm{m}(\theta; \mathbf{d}, \mathbf{x}) = \sum_{i \in C}\{d_i - \mathbf{E}_\zeta L_C(i; \mathbf{x}, \beta + \Lambda\zeta)\} \cdot W_i(\mathbf{x}, \theta) \quad (7)$$

Define $s_i(\theta; \mathbf{x}) \equiv \nabla_\theta \log P_C(i \mid \mathbf{x}, \theta)$. When $W_i(\mathbf{x}, \theta) = s_i(\theta; \mathbf{x})$, $m(\theta; \mathbf{d}, \mathbf{x})$ reduces to the score of an observation and the classical method of moments estimator coincides with the maximum likelihood estimator. However, any instrument vector $W_i(\mathbf{x}, \theta)$ whose covariance matrix with $s_i(\theta; \mathbf{x})$ is of maximum rank can be used to obtain estimators that are consistent and $N^{1/2}$ asymptotically normal, but in general less than fully efficient. A Method of Simulated Moments (MSM) estimator for MMNL is obtained by replacing $P_C(i \mid \mathbf{x}, \theta)$ in the generalized residual by the *unbiased* simulator $P_C^r(i \mid \mathbf{x}, \theta)$ and using *statistically independent* simulators (as necessary) to obtain the instrument vector $W_i(\mathbf{x}, \theta)$ for a simulator $m^r(\theta; \mathbf{d}, \mathbf{x})$ of $m(\theta; \mathbf{d}, \mathbf{x})$. The MSM estimator θ_N is a root of $\mathbf{E}_N m^r(\theta; \mathbf{d}, \mathbf{x})$. McFadden (1989, 1996) shows that under mild regularity conditions, including stochastic equicontinuity, the MSM estimator θ_N is consistent and asymptotically normal. It is not necessary for this result that r increase with N, so long as the simulators of the generalized residuals are independent or satisfy a weaker condition that is sufficient for a central limit theorem to operate across observations. The array $\Psi_N(\theta_N)^{-1}\Sigma_N(\theta_N)\Psi_N(\theta_N)^{-1}$ consistently estimates the asymptotic covariance matrix of θ_N, where

$$\Psi_N(\theta) = -\mathbf{E}_N \sum_{i \in C}\{\nabla_\theta P_C^r(i \mid \mathbf{x}, \theta)\} W_i' = -\mathbf{E}_N \sum_{i \in C} \left\{ \mathbf{E}_r \begin{bmatrix} x_{iC}(\zeta)' \\ \mathrm{vec}(\zeta x_{iC}(\zeta)) \end{bmatrix} L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \right\} W_i'$$

$$\Sigma_N(\theta) = \mathbf{E}_N \left\{ \sum_{i \in C} W_i P_C^r(i \mid \mathbf{x}, \theta) W_i' - \left[\sum_{i \in C} W_i P_C^r(i \mid \mathbf{x}, \theta) \right] \left[\sum_{i \in C} W_i P_C^r(i \mid \mathbf{x}, \theta) \right]' \right\}$$

The MSLE method is asymptotically efficient, but the computational advantages of the MSM method may offset the loss of statistical efficiency. The more highly correlated $W_i(\mathbf{x}, \theta)$ with $s_i(\theta_0; \mathbf{x})$, the more efficient the MSM estimator. An obvious candidate for $W_i(\mathbf{x}, \theta)$ is the simulated score $s_i^r(\theta; \mathbf{x})$. Large r will be needed to simulate $s_i(\theta_0; \mathbf{x})$ accurately and achieve high

efficiency. However, it is possible to obtain a computationally convenient instrument vector that is fairly highly correlated with $s_i(\theta; \mathbf{x})$, and will as a consequence yield moderately efficient MSM estimates at low computational cost. Using the approach of Talvitie (1972), make a second-order Taylor's expansion of the multinomial logit function $L_C(i; \mathbf{x}, \beta + \Lambda\zeta)$ in ζ around $\zeta = 0$, and take the expectation of this approximation with respect to ζ,

$$\mathbf{E}_\zeta L_C(i; \mathbf{x}, , \beta + \Lambda\zeta) \approx L_C(i; \mathbf{x}, \beta)\cdot\{1 + \tfrac{1}{2}\mathrm{tr}(\Lambda' Q_{ic}\Lambda)\} \tag{8}$$

where $Q_{iC} = x'_{iC}x_{iC} - \Sigma_{j\in C}L_C(j; \mathbf{x}, \beta)x'_{jC}x_{jC}$ and $x_{iC} = x_i - \Sigma_{j\in C}x_j L_C(j; \mathbf{x}, \beta)$. Because the Taylor expansion is not uniformly convergent, this is a poor approximation to the MMNL response probability itself. However, it provides an easily computed approximation to $s_i(\theta_o; \mathbf{x})$: Make a linear approximation $\log\{1 + \frac{1}{2}\mathrm{tr}(\Lambda' Q_{iC}\Lambda)\} \approx \frac{1}{2}\mathrm{tr}\{\Lambda' Q_{iC}\Lambda\}$, and take the gradient of the log of equation (8) with respect to θ, ignoring the dependence of Q_{iC} on β, to obtain $W_i(\mathbf{x}, \theta)' = [x_{iC}\mathrm{vec}(Q_{iC}\Lambda)']$. For preliminary estimation, β can be set to simple MNL coefficient estimates and Λ can be any matrix of full column rank that respects the exclusion restrictions present in the model. Limited Monte Carlo evidence suggests that use of these easily computed instruments will often yield MSM estimators with asymptotic efficiencies over 90%.[3]

4. SPECIFICATION TESTING

Because the MMNL model requires use of simulation methods, it is useful to have a specification test based solely on MNL model estimates that determine if mixing is needed. The next result describes a Lagrange Multiplier test for this purpose. This test has the pivotal property that its asymptotic distribution, under the null hypothesis that the correct specification is MNL, does not depend on the parameterization of the mixing distribution under the alternative.

Theorem 2. Consider choice from a set $\mathbf{C} = \{1, \ldots, J\}$. Let x_i be a $1 \times K$ vector of attributes of alternative i. From a random sample $n = 1, \ldots, N$, estimate the parameter α in the simple MNL model $L_C(i; \mathbf{x}, \alpha) = e^{x_i\alpha}/\Sigma_{j\in C}e^{x_j\alpha}$ using maximum likelihood; construct artificial variables

$$z_{ti} = \tfrac{1}{2}(x_{ti} - x_{tC})^2 \text{ with } x_{tC} = \sum_{j\in C} x_{tj}\cdot L_C(j; \mathbf{x}, \hat{\alpha}) \tag{9}$$

for selected components t of x_i, and use a Wald or Likelihood Ratio test for the hypothesis that the artificial variables z_{ti} should be omitted from the MNL model. This test is asymptotically equivalent to a Lagrange multiplier test of the hypothesis of no mixing against the alternative of a MMNL model $P_C(i \mid \mathbf{x}, \theta) = \int L_C(i; \mathbf{x}, \alpha)\cdot G(\mathrm{d}\alpha; \theta)$ with mixing in the selected components t of α. The degrees of freedom equals the number of artificial variables z_{ti} that are linearly independent of x.

The proof is given in the Appendix. To examine the operating characteristics of the test, we carried out two simple Monte Carlo experiments for choice among three alternatives, with random utility functions $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. The disturbances ε_i were i.i.d. Extreme Value Type I. In the first experiment, the covariates were distributed as described below:

[3] These instruments are similar to instruments for the multinomial probit model proposed independently by Ruud (1996).

Variable	Alternative 1	Alternative 2	Alternative 3
x_1	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	0	0
x_2	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	$\pm\frac{1}{2}$ w.p.	0

The parameter $\alpha_2 = 1$ under both the null and the alternative. The parameter $\alpha_1 = 0.5$ under the null hypothesis, and under the alternative $\alpha_1 = 0.5 \pm 1$ w.p. $\frac{1}{2}$. We carried out 1000 repetitions of the test procedure for a sample of size $N = 1000$ and choices generated alternately under the null hypothesis and under the alternative just described, using likelihood ratio tests for the omitted variable z_{1i}. The results are given below:

Nominal significance level	Actual significance level	Power against the alternative
10%	8·2%	15·6%
5%	5·0%	8·2%

The nominal and actual significance levels of the test agree well. The power of the test is low, and an examination of the estimated coefficients reveals that the degree of heterogeneity in tastes present in this experiment gives estimated coefficients close to their expected values. Put another way, this pattern of heterogeneity is difficult to distinguish from added extreme value noise.

In the second experiment, the covariates are distributed as shown below:

Variable	Alternative 1	Alternative 2	Alternative 3
x_1	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	0
x_2	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	$\pm\frac{1}{2}$ w.p. $\frac{1}{2}$	0

The utility function is again $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. Under the null hypothesis, $\alpha_1 = \alpha_2 = 1$, while under the alternative $(\alpha_1, \alpha_2) = (2, 0)$ w.p. $\frac{1}{2}$ and $(0, 2)$ w.p. $\frac{1}{2}$. Again, 1000 repetitions of the tests are made for $N = 1000$ under the null and the alternative; the results are given below:

Nominal significance level	Actual significance level	Power against the alternative
10%	9·7%	52·4%
5%	3·9%	39·8%

In this case where mixing is across utility functions of different variables, the test is moderately powerful. It remains the case in this example that the estimated coefficients in the MNL model without mixing are close to their expected values.

4.1. Testing the Adequacy of a Mixing Distribution

Suppose one has estimated a MMNL model in which the MNL parameters $\alpha = \beta + \Lambda\zeta$ are mixed by a base density $f(\zeta)$, and the object is to test whether *additional* mixing is needed to describe the sample. The choice probability under the alternative is

$$P_C(i \mid \mathbf{x}, \theta, \lambda) = \int \left\{ \int L_C(i; \mathbf{x}, \beta + \Lambda\zeta + \lambda^{1/2} \odot \nu) \cdot f(\zeta) \mathrm{d}\zeta \right\} \cdot h(\nu) \mathrm{d}\nu \tag{10}$$

where β is a $K \times 1$ vector, Λ is a factor loading matrix, ν is $K \times 1$ with mean zero and unit variances, λ is a $k \times 1$ vector of variances, $K - T$ of which are maintained at zero, $\lambda^{1/2}$ denotes the component-wise square root, and $\odot$ denotes the component by component direct product. The null hypothesis is that the data are generated by this model with $\lambda = 0$; i.e., a mixed MNL model with latent factors ζ determining the choice probabilities, versus the alternative that up to T additional factors, with density $h(.)$, are needed. The following theorem, proved in the Appendix, gives a Lagrange Multiplier test for this hypothesis:

Theorem 3. Suppose the base model $P_C(i \mid \mathbf{x}, \theta) = \int L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \cdot f(\zeta) \mathrm{d}\zeta$ has been estimated by MSLE, using Monte Carlo draws ζ^k from $f(\cdot)$ for $k = 1, \ldots, r$. Construct the quantities

$$x_C^k = \sum_{j \in C} x_j \cdot L_C(j; \mathbf{x}, \beta + \Lambda\zeta^k), \; z_{ti}^k = \tfrac{1}{2}(x_{ti} - x_{tC}^k)^2, \; z_{tC}^k = \sum_{j \in C} z_{tj}^k \cdot L_C(j; \mathbf{x}, \beta + \Lambda\zeta^k)$$

$$v_i = \frac{1}{r \cdot P_{\mathbf{C}}^r(i \mid \mathbf{x}, \theta)} \cdot \sum_{k=1}^{r} (x_i - x_{\mathbf{C}}^k) \cdot L_C(i; \mathbf{x}, \beta + \Lambda\zeta^k)$$

$$w_i = \frac{1}{r \cdot P_{\mathbf{C}}^r(i \mid \mathbf{x}, \theta)} \sum_{k=1}^{r} \mathrm{vec}(\zeta^k (x_i - x_C^k))' L_C(i; \mathbf{x}, \beta + \Lambda\zeta^k)$$

$$y_{ti} = \frac{1}{r \cdot P_C^r(i \mid \mathbf{x}, \theta)} \sum_{k=1}^{r} (z_{ti} - z_{tC}^k) \cdot L_C(i; \mathbf{x}, \beta + \Lambda\zeta^k)$$

where all parameters are set to the base model estimates. A regression over alternatives and observations of the integer 1 on the variables v_i, w_i, and y_{ti} for $t = 1, \ldots, T$ and an F-test for the significance of the variables in this regression is asymptotically equivalent to a Lagrange Multiplier test of the hypothesis of no additional mixing in the coefficients of x_{ti} for $t = 1, \ldots, T$.

In light of the Monte Carlo results in the base case of no mixing, one can expect this test to have relatively low power. Hence, for use as a diagnostic for model specification, one will want to err on the side of admitting too much potential heterogeneity, and use a rejection region with a large nominal significance level.

5. AN APPLICATION: DEMAND FOR ALTERNATIVE VEHICLES

The State of California suffers from air pollution generated by conventional gasoline-powered vehicles, and the State is in the process of mandating quotas for alternative-fuelled vehicles: methanol, compressed natural gas (CNG), or electric. An important policy question is consumer acceptance of these alternative vehicles, and the extent to which subsidies will be necessary to stimulate consumer demand to the levels required by the quotas. Brownstone *et al.* (1996) have carried out a conjoint analysis study of preferences between alternative vehicles. The study has 4654 respondents, each of whom was asked to choose among six alternatives. The alternatives were described in terms of the variables defined in Table I. We do not alter the variable transformations used in the original study, but note that the dependence of their specification on the price of an alternative and on income fails the quasi-convexity condition for conditional

Table I. Variable definitions

Variable	Definition
Price/log(income)	Purchase price (in thousands of dollars) divided by log(household income in thousands)
Range	Hundreds of miles vehicle can travel between refuellings/rechargings
Acceleration	Tens of seconds required to reach 30 mph from stop
Top speed	Highest attainable speed in hundreds of MPH
Pollution	Tailpipe emissions as fraction of those for new gas vehicle
Size	0=mini, 0·1=subcompact, 0·2=compact, 0·3=mid-size or large
'Big enough'	1 if household size $\geq$ and vehicle is mid or large
Luggage space	Fraction of luggage space in comparable new gas vehicle
Operating cost	Cost per mile of travel (tens of cents): home recharging for electric vehicle, station refuelling otherwise
Station availability	Fraction of stations that can refuel/recharge vehicle
Sports utility vehicle	1 if sports utility vehicle, 0 otherwise
Sports car	1 if sports car
Station wagon	1 if station wagon
Truck	1 if truck
Van	1 if van
EV	1 if electric vehicle (EV)
Commute <5 & EV	1 if electric vehicle and commute <5 miles/day
College & EV	1 if electric vehicle and some college education
CNG	1 if compressed natural gas (CNG) vehicle
Methanol	1 if methanol vehicle
College & Methanol	1 if methanol vehicle and some college education

indirect utility that comes from economic consumer theory. An experimental design was used to select the offerings of six alternatives from 120 possible profiles, distinguished by four fuels (gasoline, methanol, CNG, electric), five sizes (mini, subcompact, compact, midsize, large), and six body types (regular car, sports car, truck, van, station wagon, sports utility vehicle).

Table II gives a MMNL model estimated by Brownstone and Train (1999). This model includes four random effects, associated with the following variables: Dummy for non-EV, Dummy for non-CNG, Size, and Luggage Space. The segment of the table headed 'Variables' gives estimates of the β parameters, and the segment headed 'Random Effects' gives the factor loading Λ on standard normal factors, with an independent factor for each of the random effects above. Then, the coefficients are estimates of the standard deviations of these random effects. The estimation uses 250 replications per observation, and MSLE. The parameter estimates show strong random effects, with magnitudes large enough to suggest that they are capturing correlation structure in unobservables in addition to variation in tastes. The variables and random effects included in this model are the result of a classical selection procedure that estimated alternative MMNL models and used a likelihood ratio test to select from them. A likelihood ratio test at the 5% level shows that this model fits significantly better than a simple MNL model (given in Table III). The table gives estimates of the standard errors of the coefficients for 250 replications, and also for 50 replications. The columns headed 'Asymptotic' give standard errors using $\Delta_N(\theta_N)^{-1}$. As noted in Section 3, while this estimator is consistent, for moderate r it can underestimate covariances and lead to the perverse conclusion that standard errors increase when the number of simulation draws rises. The columns headed 'Robust' give

standard errors using the recommended covariance matrix estimator $\Gamma_N(\theta_N)^{-1}\Delta_N(\theta_N)\Gamma_N(\theta_N)^{-1}$. The 'Robust' standard errors fall with number of repetitions, as expected. In general, using the 'Asymptotic' covariance formula with 250 replications results in a 10–20% underestimate of standard errors of coefficients, compared to the 'Robust' formula.

Table III, Model 1, is a simple MNL model $L_C(i; \mathbf{x}, \beta)$ fitted to the data; these estimates are taken from Brownstone and Train (1999). Model 2 adds the artificial variables defined in Theorem 2; i.e. given the base MNL model $L_C(i; \mathbf{x}, \beta)$ and $x_C = \Sigma_{j\in C} x_j \cdot L_C(j; \mathbf{x}, \beta)$, with β set equal to its MNL estimator, define the artificial variables $z_{ti} = \frac{1}{2}(x_{ti} - x_{tC})^2$ for variables t where heterogeneity is suspected, and estimate the MNL model with the original x variables and the additional artificial variables. The list of artificial variables may include variables t which have the coefficient β_t constrained to zero in the base MNL model; these are interpreted as pure random effects. A likelihood ratio test at the 5% significance level rejects the null hypothesis of no mixing. The individual T-statistics for the artificial variables are not necessarily a reliable guide to

Table II. Mixed logit for alternative-fuelled vehicle choice

	Parameter estimates	Standard error: 250 replications		Standard error: 50 replications	
		Asymptotic	Robust	Asymptotic	Robust
Variables					
Price/log(income)	−0·264	0·0435	0·0452	0·0412	0·0525
Range	0·517	0·0581	0·0685	0·0511	0·1022
Acceleration	−1·062	0·1859	0·1990	0·1738	0·2519
Top speed	0·307	0·1150	0·1184	0·1131	0·1188
Pollution	−0·608	0·1392	0·1420	0·1357	0·1546
Size	1·435	0·5082	0·4991	0·4945	0·5156
'Big enough'	0·224	0·1126	0·1166	0·1113	0·1220
Luggage space	1·702	0·4822	0·5854	0·4314	0·8971
Operating cost	−1·224	0·1593	0·2069	0·1393	0·2998
Station availability	0·615	0·1452	0·1536	0·1410	0·1757
Sports utility vehicle	0·901	0·1484	0·1486	0·1482	0·1493
Sports car	0·700	0·1625	0·1513	0·1626	0·1518
Station wagon	−1·500	0·0674	0·0645	0·0674	0·0659
Truck	−1·086	0·0556	0·0520	0·0555	0·0556
Van	−0·816	0·0558	0·0468	0·0557	0·0471
EV	−1·032	0·4249	0·5022	0·3777	0·6035
Commute <5 & EV	0·372	0·1660	0·1763	0·1608	0·1927
College & EV	0·766	0·2182	0·2374	0·2073	0·2796
CNG	0·626	0·1482	0·1670	0·1391	0·2139
Methanol	0·415	0·1464	0·1474	0·1440	0·1534
College & Methanol	0·313	0·1243	0·1256	0·1223	0·1308
Random effects					
Non-EV	2·464	0·5414	0·7184	0·4428	1·0252
Non-CNG	1·072	0·3773	0·4109	0·2781	0·5711
Size	7·455	1·8194	2·0408	1·5538	2·4734
Luggage Space	5·994	1·2483	1·6617	1·0483	2·7719
Log likelihood	−7375·34				

Note: Parameter estimates are from Brownstone and Train (1999); standard error estimates are from this study.

the location of significant mixing, due to lack of independence, and to the possibility of correlation across alternatives in unobserved attributes. However, the results (based on *T*-statistics exceeding one in magnitude) suggest that there may be taste variation in the following variable coefficients: Non-EV, Non-CNG, Size, Luggage space, Operating Cost, and Station Availability. The first four of these were included in the Brownstone–Train model in Table II; the last two are additional factors where mixing may be present. Our specification testing procedure

Table III. Multinomial logit model

Variables	Model 1		Model 2	
	Parameter estimate	SE	Parameter estimate	SE
Price/log(income)	−0·185	0·027	−0·4240	0·0298
Range	0·350	0·027	0·5036	0·0447
Acceleration	−0·716	0·111	−0·9771	0·1263
Top speed	0·261	0·080	0·3592	0·0814
Pollution	−0·444	0·100	−0·6567	0·1161
Size	0·935	0·311	1·4179	0·3430
'Big enough'	0·143	0·076	0·2248	0·0845
Luggage space	0·501	0·188	1·0161	0·2574
Operating cost	−0·768	0·073	−1·1447	0·0897
Station availability	0·413	0·097	0·6350	0·1074
Sports utility vehicle	0·820	0·144	0·8806	0·1458
Sports car	0·637	0·156	0·6869	0·1580
Station wagon	−1·437	0·065	−1·5229	0·0663
Truck	−1·017	0·055	−1·0776	0·0551
Van	−0·799	0·053	−0·8272	0·0542
EV	−0·179	0·169	−0·6979	0·2384
Commute <5 & EV	0·198	0·082	0·3102	0·0840
College & EV	0·443	0·108	0·6863	0·1145
CNG	0·345	0·091	0·4216	0·1056
Methanol	0·313	0·103	0·4886	0·1105
College & Methanol	0·228	0·089	0·3070	0·0903
Artificial variables				
Price/log(income)			0·0019	0·0927
Range			−0·0349	0·0551
Acceleration			−1·3728	2·1388
Top speed			−0·2071	0·6383
Pollution			0·0977	0·6764
Size			21·5773*	9·5000
'Big enough'			0·2837	0·3832
Luggage space			3·8731*	3·4638
Operating cost			4·2245*	0·8369
Station availability			0·6741*	0·3781
EV			2·3476*	0·5704
CNG			1·2364*	0·4798
Log likelihood	−7391·83		−7356·61	

Notes: Model 1 is from Brownstone and Train (1999).
*denotes the artificial variables with $|T|>1$.

Table IV. Mixed multinomial logit model

	Parameter estimates	SE
Variables		
Price/log(income)	−0·3622	0·0669
Range	0·6753	0·0965
Acceleration	−1·2688	0·2591
Top speed	0·4027	0·1553
Pollution	−0·7929	0·1980
Size	1·7351	0·6694
'Big enough'	0·2695	0·1468
Luggage space	2·2631	0·6426
Operating cost	−1·8056	0·2912
Station availability	0·7029	0·1896
Sports utility vehicle	0·9234	0·1498
Sports car	0·7270	0·1645
Station wagon	−1·5246	0·0681
Truck	−1·1195	0·0559
Van	−0·8191	0·0564
EV	−1·5733	0·5819
Commute <5 & EV	0·4793	0·2242
College & EV	1·0534	0·3114
CNG	0·7709	0·2018
Methanol	0·5435	0·1922
College & Methanol	0·3849	0·1542
Random effects		
Non-EV	3·3802	0·7647
Non-CNG	1·1042	0·4990
Size	8·0788	2·7021
Luggage space	7·6220	1·7153
Operating cost	4·4532	0·8014
Station availability	1·3987	0·5730
Log likelihood	−7358·93	

is easier and quicker than the Brownstone–Train method. All the factors identified in their search were picked up by our procedure, as were some additional candidates.

Table IV gives a MMNL model which includes the six random effects identified as possibly significant by the artificial variable test in Table III, using T-statistics greater than one in magnitude as the selection criterion. The MMNL estimates show that there is significant mixing in each of these factors. Likelihood ratio tests show that this model is a significant improvement on the model in Table II. Further exploration with additional factors in the MMNL model finds that there are several factor combinations that will fit as well or marginally better than the model in Table IV, and that some of these combinations will place weight on factors that were excluded by the artificial variable selection procedure, and will lower the significance of some of the factors previously included. These results reflect the inherent difficulty of identifying the factor structure of unobserved utility from observed data on discrete choices, but may also indicate more conventional specification issues such as omitted observed variables or interactions.

6. CONCLUSIONS

This paper has established that MMNL models, estimated using MSLE or MSM, provide a flexible and computationally practical econometric method for economic discrete choice that is postulated to come from utility maximization. First, a general approximation property is established. Second, estimation of parametric MMNL models by MSLE or MSM is shown to provide estimates with good statistical properties, and easily computed fairly efficient instruments are provided for MSM. Third, simply computed specification tests are developed that allow one to test for the presence of mixing, or for the presence of omitted mixing factors. Finally, an application to the demand for alternatively-fuelled vehicles shows that the method can detect and estimate significant mixing effects which can have a strong effect on the pattern of substitutability across alternatives.

APPENDIX: PROOFS OF THEOREMS

Lemma 1. Suppose consumers with tastes defined by points s in a compact topological space $\mathbf{S}$ have preferences over objects z in a compact topological space $\mathbf{Z}$, with $z' \succeq_s z''$ meaning z' is at least as good as z'' for a consumer with tastes s. Suppose $\succeq_s$ is complete and transitive, and has the continuity property that if a sequence of triples (z'^k, z''^k, s^k) converges to a limit (z'^0, z''^0, s^0) and satisfies $z'^k \succeq_{s^k} z''^k$, then $z'^0 \succeq_s z''^0$. Then there exists a utility function $U(z, s)$, continuous in its arguments, that represents $\succeq_s$ for $s \in \mathbf{S}$.

Proof: A standard construction for fixed s due to Rader and Debreu defines a utility function $U(z, s)$ which is continuous in z for each s; see Barten and Bohm (1982, pp. 388–390). But the level sets $\{(z', s') \mid U(z', s') \geqslant U(z, s)\}$ and $\{(z', s') \mid U(z', s') \leqslant U(z, s)\}$ are then closed by the hypothesized continuity property, implying that U is continuous on $\mathbf{Z} \times \mathbf{S}$.

Lemma 2. Consider a random variable X with CDF F. Define $F(x^-) = \lim_{\varepsilon \searrow 0} F(x - \varepsilon)$ and the right-continuous inverse $F^{-1}(p) = \sup\{x \in \mathbb{R} \mid F(x) \leqslant p\}$ for $p \in (0, 1)$. Define the random variable $Z = U \cdot F(X) + (1 - U) \cdot F(X^-)$, where U is a uniformly distributed random variable on [0,1] that is independent of X. Define $X^* = F^{-1}(Z)$. Then Z is uniformly distributed on (0,1), the function F^{-1} is almost surely continuous, and $X^* = X$ almost surely. If, in addition, F is strictly increasing, then F^{-1} is continuous and $X^* \equiv X$.

Proof: Define $\mathbf{A} = \{x \in \mathbb{R} | F(x + \varepsilon) = F(x) \text{ for some } \varepsilon > 0\}$, and $\mathbf{B} = F(\mathbf{A})$. The set $\mathbf{A}$ is a countable union of disjoint half-open intervals of the form $[a,b)$ with $F(a) = F(b^-)$, so that $\mathbf{A}$ occurs with probability zero. For $c \in (0, 1)$, let $x = F^{-1}(c)$. Then, for all $\varepsilon > 0$, $F(x - \varepsilon) \leqslant F(x^-) \leqslant c \leqslant F(x) < F(x + \varepsilon)$. Hence, the event $\{Z \leqslant c\}$ occurs if and only if one of the disjoint events $\{X < x\}$ or $\{X = x \;\&\; F(x^-) + U \cdot [F(x) - F(x^-)] < c\}$ occurs. But $P(X < x) = F(x^-)$ and $P(X = x \;\&\; F(x^-) + U \cdot [F(x) - F(x^-)] < c) = c - F(x^-)$, implying $P(Z \leqslant c) = c$. Thus, Z is uniformly distributed on (0,1). If $x \notin \mathbf{A}$ and $F(x^-) \leqslant c \leqslant F(x)$, then $F^{-1}(c) = x$. Then, in the event $X \notin \mathbf{A}$, which occurs with probability one, one has $X = X^* \equiv F(Z)$. Finally, $F^{-1}(p)$ is a monotone right-continuous function that is also left-continuous except at jumps when $p \in \mathbf{B}$, a countable set that contains Z with probability zero. Then, F^{-1} is almost surely continuous. If F is strictly increasing, then $\mathbf{A}$ and $\mathbf{B}$ are empty. ■

Let $(\Omega, \mathcal{W}, \pi)$ denote a probability space, $\mathbf{T}$ a subset of $\mathbb{R}^m$, and $X : \Omega \times \mathbf{T} \to \mathbb{R}^n$ a random field, measurable with respect to $\mathcal{W}$ for each $t \in \mathbf{T}$, and almost surely measurable on $\mathbf{T}$. We say that X admits a *coordinate conditional probability structure* at t if there exist conditional CDF functions

$F_i(x_i \mid x_1, \ldots, x_{i-1}, t)$, measurable in their arguments, such that for all $y \in \mathbb{R}^n$, the CDF of $X(t)$ can be written

$$F(y, t) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \cdots \int_{-\infty}^{y_n} F_1(\mathrm{d}x_1 \mid t) \cdot F_2(\mathrm{d}x_2 \mid x_1, t) \cdot \ldots \cdot F_n(\mathrm{d}x_n \mid x_1, \ldots, x_{n-1}, t)$$

This condition follows from Fubini's theorem if $F(x, t)$ has a density $f(x, t)$, and will hold automatically if F is built up from conditional CDF functions. It can fail in pathological cases; see Billingsley (1986, p. 458).

Lemma 3. Suppose X is a random field from $\mathbf{T} \subseteq \mathbb{R}^m$ into $\mathbb{R}^n$ that admits a coordinate conditional probability structure for each $t \in \mathbf{T}$. Then there exists a uniformly distributed random field $Z : \Omega \times \mathbf{T} \to [0, 1]^n$ and measurable functions $h_i : [0, 1]^i \times \mathbf{T} \to \mathbb{R}$ for $i = 1, \ldots, n$ such that $X_i(t) = h_i(Z_1(t), \ldots, Z_i(t), t)$ almost surely. If in addition, there is a rectangle $\mathbf{X} \subseteq \mathbb{R}^n$ such that $X(t)$ has a density $f(x, t)$ that is continuous in its arguments and strictly positive on the interior of $\mathbf{X}$, with $F(\mathbf{X}, t) = 1$, then the h_i are continuous in their arguments and Z is a continuous random field.

Proof: For $n = 1$, Lemma 2 gives the result for each $t \in \mathbf{T}$ that $X_1(t) = h_1(z_1(t), t)$ almost surely. Proceed by induction. Suppose the result has been established for $n - 1$, with $Z_i(t)$ independently distributed uniformly on [0,1] for $i = 1, \ldots, n - 1$ and $X_i(t) = h_i(Z_1(t), \ldots, Z_i(t), t)$ almost surely. Apply Lemma 2 to the random variable $X_n(t)$ with measurable conditional distribution function $F_n(x_n \mid h_1(Z_1(t), t), \ldots, h_{n-1}(Z_1(t), \ldots, Z_{n-1}(t), t), t)$. This yields a function satisfying $X_n(t) = h_n(Z_1(t), \ldots, Z_n(t), t)$ almost surely. This completes the induction step. Finally, since X is a random field, almost sure continuity assures that the pointwise construction for each $t \in \mathbf{T}$ is measurable in t.

When X has a positive continuous density $f(x, t)$ on the interior of the rectangle $\mathbf{X}$, the conditional CDF $F_i(x_i \mid x_1, \ldots, x_{i-1}, t)$ is strictly increasing in x_i and continuous in all its arguments, so that an implicit function theorem implies that the inverse function $x_i = F_i^{-1}(z_i \mid x_1, \ldots, x_{i-1}, t)$ is continuous in all its arguments, giving the last result of the lemma. ■

Lemma 4. A necessary and sufficient condition for a random utility function over J alternatives with a multivariate CDF $F(u)$ to have a representation $U = W + \varepsilon$, where the components of ε are independent of W and independent identically distributed Extreme Value I, is that $F(-\log(t_1), \ldots, -\log(t_J))$ be a multivariate Laplace Transform; i.e. F treated as a function of $(t_1, \ldots, t_J)$ is analytic and has derivatives satisfying the sign condition

$$(-1)^{n_1+\cdots+n_J} \cdot \partial^{n_1+\cdots+n_J} F(-\log(t_1), \ldots, -\log(t_J)) / \partial^{n_1} t_1 \ldots \partial^{n_J} t_J \geqslant 0$$

Proof: Suppose U has the representation. Then, F must satisfy the convolution formula

$$F(u_1, \ldots, u_J) = \int_{-\infty}^{+\infty} \exp(-(e^{-u_1+w_1} + \cdots + e^{-u_J+w_J})) \cdot G(\mathrm{d}w)$$

where G is the CDF of W. Make the transformations $t_k = \exp(-u_k)$ and $V_k = \exp(W_k)$, and let H be the CDF of V. Then,

$$F(-\log(t_1), \ldots, -\log(t_J)) = \int_0^{+\infty} \exp(-(t_1 v_1 + \cdots + t_J v_J)) \cdot H(\mathrm{d}v)$$

This is a multivariate Laplace Transform. Conversely, if $F(-\log(t_1), \ldots, -\log(t_J))$ is a multivariate Laplace Transform, then it satisfies $F(-\infty) = 0$, $F(+\infty) = 1$, and from the derivative property of Laplace Transforms, $\partial^J F / \partial u_1 \ldots \partial u_J \geqslant 0$, so that it is a multivariate CDF; see Feller (1966, p. 415). ■

Lemma 5. If $P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_\zeta L_C(i; \mathbf{x}, \beta + \Lambda\zeta)$ is given by equation (5), then

$$\nabla_{\beta\beta'} \log P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_{\zeta \mid i}\left\{x_{iC}(\zeta)' x_{iC}(\zeta) - \Sigma_{j \in C} x_{jC}(\zeta)' x_{jC}(\zeta) L_C(j; \mathbf{x}, \beta + \Lambda\zeta)\right\} - \left\{\nabla_\beta \log P_C(i \mid \mathbf{x}, \theta)\right\}\left\{\nabla_\beta \log P_C(i \mid \mathbf{x}, \theta)\right\}'$$

$$\nabla_{\gamma\beta'} \log P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_{\zeta \mid i}\left\{\mathrm{vec}(\zeta x_{iC}(\zeta)) x_{iC}(\zeta) - \Sigma_{j \in C} \mathrm{vec}(\zeta x_{jC}(\zeta)) x_{jC}(\zeta) L_C(j; \mathbf{x}, \beta + \Lambda\zeta)\right\} - \left\{\nabla_\gamma \log P_C(i \mid \mathbf{x}, \theta)\right\}\left\{\nabla_\beta \log P_C(i \mid \mathbf{x}, \theta)\right\}'$$

$$\nabla_{\gamma\gamma'} \log P_C(i \mid \mathbf{x}, \theta) = \mathbf{E}_{\zeta \mid i}\left\{\mathrm{vec}(\zeta x_{iC}(\zeta))\mathrm{vec}(\zeta x_{jC}(\lambda))' - \Sigma_{j \in C} \mathrm{vec}(\zeta x_{jC}(\lambda))\mathrm{vec}(\zeta x_{jC}(\lambda))' L_C(j; \mathbf{x}, \beta + \Lambda\zeta)\right\} - \left\{\nabla_\gamma \log P_C(i \mid \mathbf{x}, \theta)\right\}\left\{\nabla_\gamma \log P_C(i \mid \mathbf{x}, \theta)\right\}'$$

Proof: Direct computation. ■

Proof of Theorem 1: Consider the random utility model $U^*(z, s, \varepsilon(\omega, z), \upsilon(\omega, s))$, where $\varepsilon(\omega, z) \in [0, 1]^p$ and $\upsilon(\omega, s) \in [0, 1]^r$ are uniformly distributed continuous random fields defined for ω in a fundamental probability space $(\Omega, \mathcal{W}, \pi)$. For $(z', z'') \in \mathcal{C}_2$ and $s \in \mathbf{S}$, define the set

$$\mathbf{A}_k(z', z'', s) = \{\omega \in \Omega \mid \left|U^*(z', s, \varepsilon(\omega, z'), \upsilon(\omega, s)) - U^*(z'', s, \varepsilon(\omega, z''), \upsilon(\omega, s))\right| \geqslant 5/k\}$$

The continuity of U^* and the measurability of the random fields implies that $\mathbf{A}_k(z', z'', s)$ is measurable. This set is monotone increasing as $k \to \infty$ to the set of ω for which the alternatives $(z', z'') \in \mathcal{C}_2$ are not tied. By hypothesis, this set has probability one, implying that there exists $k = k(z', z'', s)$ such that $\pi(\mathbf{A}_k(z', z'', s)) > 1 - \eta/4J^*$.

The uniform continuity of U^* on $\mathbf{Z} \times \mathbf{S} \times [0, 1]^p \times [0, 1]^r$ implies that given $k(z', z'', s)$, there exists $\delta(z', z'', s) > 0$ such that in a neighbourhood of this radius U^* varies by less than $1/k(z', z'', s)$. The almost certain continuity of $\varepsilon(\omega, z)$ and $\upsilon(\omega, s)$ imply that the set

$$\mathbf{B}_m(z', s) = \{\omega \in \Omega \mid \sup_{|z^* - z'| < 1/m} |\varepsilon(\omega, z^*) - \varepsilon(\omega, z')| + \sup_{|s^* - s| < 1/m} |\upsilon(\omega, s^*) - \upsilon(\omega, s)| < \delta(z', z'', s)\}$$

and the corresponding set $\mathbf{B}_m(z'', s)$, are monotone increasing as $m \to \infty$ to limiting sets that occur with probability one. Then there exists $m = m(z', z'', s)$ such that $\pi(\mathbf{B}_m(z', s)) > 1 - \eta/4J^*$ and $\pi(\mathbf{B}_m(z'', s)) > 1 - \eta/4J^*$. Therefore with probability at least $1 - 3\eta/4J^*$,

$\omega \in \mathbf{A}_k(z', z'', s) \cap \mathbf{B}_m(z', s) \cap \mathbf{B}_m(z'', s)$, implying that for $(z^{*\prime}, z^{*\prime\prime}, s^*)$ in an open neighbourhood of $(z', z'', s) \in \mathcal{C}_2 \times \mathbf{S}$ of radius $\delta(z', z'', s)$, one has

$$| U^*(z^{*\prime}, s^*, \varepsilon(\omega, z^{*\prime}), \upsilon(\omega, s^*)) - U^*(z^{*\prime\prime}, s^*, \varepsilon(\omega, z^{*\prime\prime}), \upsilon(\omega, s^*)) | \geqslant 1/k$$

These neighbourhoods cover the compact set $\mathcal{C}_2 \times \mathbf{S}$. Therefore, there exists a finite subcovering. Let k^* be the larger of $-\log(\eta/4J^*)$ and the maximum value of $k(z', z'', s)$ for the centres of the finite subcover. We have now established that each point $(z^{*\prime}, z^{*\prime\prime}, s^*) \in \mathcal{C}_2 \times \mathbf{S}$ falls in some neighbourhood in a finite cover with centre (z', z'', s), and satisfies $| U^*(z^{*\prime}, s^*, \varepsilon(\omega, z^{*\prime}), \upsilon(\omega, s^*)) - U^*(z^{*\prime\prime}, s^*, \varepsilon(\omega, z^{*\prime\prime}), \upsilon(\omega, s^*)) | \geqslant 3/k(z', z'', s) \geqslant 3/k^*$ on a set of ω that occurs with probability at least $1 - 3\eta/4J^*$.

The continuous function U^* has a Bernstein–Weierstrauss polynomial approximation U^{*k} on $\mathbf{Z} \times \mathbf{S} \times [0, 1]^{p+r}$ that satisfies $| U^* - U^{*k} | \leqslant 1/k^*$. Consider a choice set $\mathbf{C} = \{z_1, \ldots, z_J\} \in \mathcal{C}^*$ and let $\mathbf{z} = (z_1, \ldots, z_J)$. Form $U^k(z_i, s, \varepsilon(z_i), \upsilon(s)) = U^{*k}(z_i, s, \varepsilon(z_i), \upsilon(s)) + \nu_i/k^{*2}$, where the ν_i are i.i.d. Extreme Value Type I random variables. Consider the event of a preference reversal between U^* and U^k for a pair $(z_i, z_j) \subseteq \mathbf{C}$ and $s \in \mathbf{S}$; i.e. the set of $\omega \in \Omega$ and $\nu \in \mathbb{R}^2$ such that $U^*(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) > U^*(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s))$ and $U^k(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) < U^k(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s))$. If $| U^*(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) - U^*(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s)) | > 3/k^*$, then

$$\begin{aligned}
0 &> U^k(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) - U^k(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s)) \\
&= U^{*k}(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) - U^{*k}(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s)) + (\nu_i - \nu_j)/k^{*2} \\
&\geqslant U^*(z_i, s, \varepsilon(\omega, z_i), \upsilon(\omega, s)) - U^*(z_j, s, \varepsilon(\omega, z_j), \upsilon(\omega, s)) - 2/k^* + (\nu_i - \nu_j)/k^{*2} \\
&\geqslant 1/k^* + (\nu_i - \nu_j)/k^{*2}
\end{aligned}$$

and hence $\nu' - \nu'' < -1/k^*$. The probability of this last event is $(1 + e^{k^*})^{-1} < \eta/4J^*$, and from the previous argument the probability that the conditioning event does not occur is at most $3\eta/4J^*$. Then the probability of a preference reversal at (z_i, z_j, s) is at most η/J^*. Therefore, the probability of the event that the alternative in $\mathbf{C}$ that maximizes U^k differs from the alternative that maximizes U^* is at most $\eta J/J^* \leqslant \eta$.

Write the polynomial approximation U^{k} in the form $U^k(z, s, \varepsilon(z), \upsilon(s)) = x(z, s)\cdot\alpha(z, s) + \nu/k^{*2}$, where $x(z, s)$ is a vector of the z and s components of the terms in the polynomial and $\alpha(z, s)$ is a vector of the corresponding $\varepsilon(z)$ and $\upsilon(s)$ components. Finally, for a choice set $\mathbf{C} = \{z_1, \ldots, z_J\} \in \mathcal{C}^*$, define $\alpha = (\alpha(z_1, s), \ldots, \alpha(z_J, s))$ and $x_i = (0, \ldots, 0, x(z_i, s), 0, \ldots, 0)$, so that $U^k(z_i, s, \varepsilon_i) = x_i\cdot\alpha + \nu_i/k^{*2}$. This is a MMNL model of the form of equation (1), and the construction guarantees that with probability at least $1 - \eta$, U^* and U^k are maximized at the same alternative in $\mathbf{C}$. Therefore, $P^*_C(i \mid \mathbf{z}, s)$ from U^* and $P_C(i \mid \mathbf{x}, \theta)$ from U^k differ by at most η. ■

Proof of Theorem 2: Write the MMNL model as $P_C(i \mid \mathbf{x}, \theta, \lambda) = \int L_C(i; \mathbf{x}, \beta + \lambda^{1/2} \odot \zeta)\cdot G(\mathrm{d}\zeta; \theta)$, where β and λ are vectors of parameters, $\lambda^{1/2}$ is the vector of square roots of the components of λ, ζ is a vector of random variables that has mean zero, component variances of one, and full rank over the specified components r, and $\lambda^{1/2} \odot \zeta$ denotes the component-by-component direct product. The parameterization $\lambda^{1/2}$ is chosen to circumvent the problem that a natural parameterization in terms of the standard deviations of the mixing density leads to a score that is identically zero under the null, as in Lee and Chesher (1986), McFadden (1987), and Newey and McFadden (1994). Then

$$\nabla_\beta P_C(i\,|\,\mathbf{x},\theta,\lambda)=\int L_C(i;\mathbf{x},\beta+\lambda^{1/2}\odot\zeta)\cdot(x_i-x_C)\cdot G(\mathrm{d}\zeta;\theta)$$

$$\text{with } x_C=\Sigma_{j\in C}x_j\cdot L_C(j;\mathbf{x},\beta+\lambda^{1/2}\odot\zeta)$$

$$\nabla_{\lambda_t} P_C(i\,|\,\mathbf{x},\theta,\lambda)=\tfrac{1}{2}\cdot\lambda^{-1/2}\cdot\int L_C(i;\mathbf{x},\beta+\lambda^{1/2}\odot\zeta)\cdot(x_i-x_C)\cdot\zeta_t\cdot G(\mathrm{d}\zeta;\theta)$$

Taking the limit as the $\lambda_t\to 0$, and using L'Hôpital's rule on $\nabla_{\lambda_t}P_C(i\,|\,\mathbf{x},\theta,\lambda)$, one obtains

$$\nabla_\beta P_C(i\,|\,\mathbf{x},\theta,\lambda)=L_C(i;\mathbf{x},\beta_e)\cdot(x_i-x_C)\text{ and }\nabla_{\lambda_t}\log P_C(i\,|\,\mathbf{x},\theta)=L_C(i;\mathbf{x},\beta_e)\cdot(z_{ti}-z_{tC})$$

where $z_{ti}=\frac{1}{2}(x_{ti}-x_{tC})^2$ and $z_{tC}=\Sigma_{j\in C}z_{tj}\cdot L_C(j;\mathbf{x},\hat{\alpha})$. The sample mean of $\nabla_\beta\log P_C(i\,|\,\mathbf{x},\theta)$ is zero at the maximum likelihood estimator β_e of the simple MNL model, and the Lagrange Multiplier statistic tests whether the vector of sample means of $\nabla_{\lambda_t}\log P_C(i\,|\,\mathbf{x},\theta)$ for the selected t are zero. As in McFadden (1987), this test is equivalent to a Lagrange Multiplier test for the null hypothesis that the variables z_{ri} have zero coefficients in the MNL model, and thus asymptotically equivalent to a Likelihood Ratio or Wald test for this hypothesis. ■

Proof of Theorem 3: Consider $P_C(i\,|\,\mathbf{x},\theta)=\int\{\int L_C(i;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\nu\}H(\mathrm{d}\nu)$. Differentiating, $\nabla_\beta P_C(i\,|\,\mathbf{x},\theta)=\int\{\int(x_i-x_C)L_C(i;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\nu\}H(\mathrm{d}\nu)$, $\nabla_{\Lambda'}P_C(i\,|\,\mathbf{x},\theta)=\int\{\int\zeta(x_i-x_C)L_C(i;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\nu\}H(\mathrm{d}\nu)$, and $\nabla_{\lambda_t}P_C(i\,|\,\mathbf{x},\theta)=\frac{1}{2}\lambda_t^{-1/2}\int\{\int(x_{ti}-x_{tC})L_C(i;\mathbf{x},\beta+\Lambda\zeta\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\zeta\}\cdot\nu_t\cdot H(\mathrm{d}\nu)$, with $x_C\equiv x_C(\zeta)=\Sigma_{j\in C}x_jL_C(j;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)$. To evaluate the last derivative under the null, use L'Hôpital's rule. The derivative of $2\lambda_t^{1/2}\nabla_{\lambda_t}P_C(i\,|\,\mathbf{x},\theta)$ with respect to λ_t is

$$\tfrac{1}{2}\lambda_t^{-1/2}\int\left\{\int(x_{ti}-x_{tC})^2L_C(i;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\zeta\right\}\cdot\nu_t^2\cdot H(\mathrm{d}\nu)$$

$$-\tfrac{1}{2}\lambda_t^{-1/2}\int\left\{\int\sum\nolimits_{j\in C}x_{tj}(x_{tj}-x_{tC})L_C(j;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)L_C(i;\mathbf{x},\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\zeta\right\}\cdot\nu_t^2\cdot H(\mathrm{d}\nu)$$

$$=\tfrac{1}{2}\lambda_t^{-1/2}\int\left\{\int(z_{ti}-z_{tC})^2L_C(i;\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)\cdot f(\zeta)\mathrm{d}\zeta\right\}\cdot\nu_t^2\cdot H(\mathrm{d}\nu),$$

with $z_{ti}\equiv z_{ti}(\zeta)=(x_{ti}-x_{tC}(\zeta))^2/2$ and $z_{tC}\equiv z_{tC}(\zeta)=\Sigma_{j\in C}z_{tj}(\zeta)\cdot L_C(j\,|\,\mathbf{x},\beta+\Lambda\zeta+\lambda^{1/2}\odot\nu)$. Hence, at $\delta=0$,

$$P_C(i\,|\,\mathbf{x},\theta)=\int L_C(i;\mathbf{x},\beta+\Lambda\zeta)\cdot f(\zeta)\mathrm{d}\zeta,$$

$$\nabla_\beta P_C(i\,|\,\mathbf{x},\theta)=\int(x_i-x_C(\zeta))L_C(i;\mathbf{x},\beta+\Lambda\zeta)\cdot f(\zeta)\mathrm{d}\zeta,$$

$$\nabla_{\Lambda'}P_C(i\,|\,\mathbf{x},\theta)=\int\zeta(x_i-x_C(\zeta))L_C(i;\mathbf{x},\beta+\Lambda\zeta)\cdot f(\zeta)\mathrm{d}\zeta,\text{ and}$$

$$\nabla_{\lambda_t}P_C(i\,|\,\mathbf{x},\theta)=\int(z_{ti}(\zeta)-z_{tC}(\zeta))L_C(i;\mathbf{x},\beta+\Lambda\zeta)\cdot f(\zeta)\mathrm{d}\zeta.$$

For comparison, suppose one had the base model in variables x and wanted to test whether additional variables z_{ti} belong in the model. The model under the alternative is

$P_C(i \mid \mathbf{x}, \mathbf{z}, \theta, \alpha) = \int L_C(i; \mathbf{x}, \mathbf{z}(\zeta), \beta + \Lambda\zeta, \alpha) \cdot f(\zeta) \mathrm{d}\zeta$. The derivatives under the null hypothesis $\alpha = 0$ are the same as before for $\nabla_\beta P_C(i \mid \mathbf{x}, \mathbf{z}, \theta, \alpha)$ and for $\nabla_{\Lambda'} P_C(i \mid \mathbf{x}, \mathbf{z}, \theta, \alpha)$. Finally, $\nabla_\alpha P_C(i \mid \mathbf{x}, \mathbf{z}, \theta, \alpha) = \int (z_i(\zeta) - z_C(\zeta)) L_C(i; \mathbf{x}, \beta + \Lambda\zeta) \cdot f(\zeta) \mathrm{d}\zeta$, also as before. Therefore, a LM test for the hypothesis $\lambda = 0$ is equivalent to a LM test for the hypothesis $\alpha = 0$ for the auxiliary variables $z_i(\zeta)$. This test is readily computed by first estimating the base model using a simulation procedure with specified starting seeds, then regressing (over observations and alternatives) the integer 1 on the scores $\nabla_\beta \log P_C^r(i \mid \mathbf{x}, \theta)$, $\mathrm{vec}(\nabla_{\Lambda'} \log P_C^r(i \mid \mathbf{x}, \theta))$, and $\nabla_{\lambda_t} \log P_C^r(i \mid \mathbf{x}, \theta)$ for $t = 1, \ldots T$, and testing whether the sum of squared residuals is significant according to a chi-square distribution with T degrees of freedom. ■

ACKNOWLEDGEMENTS

We are indebted to the E. Morris Cox fund for research support, and to Moshe Ben-Akiva, David Brownstone, Denis Bolduc, Andre de Palma, and Paul Ruud for useful comments. This paper was first presented at the University of Paris X in June 1997.

REFERENCES

Barten A, Bohm V. 1982. Consumer theory. In *Handbook of Mathematical Economics*, Vol. II, Arrow K, Intriligator M (eds). North-Holland: Amsterdam.

Beggs J. 1988. A simple model for heterogeneity in binary logit models. *Economics Letters* **27**: 245–249.

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

Billingsley P. 1986. *Probability and Measure*. Wiley: New York.

Bolduc D, Fortin B, Gordon S. 1996. Multinomial probit estimation of spatially interdependent choices. *International Regional Science Review*, forthcoming.

Boyd J, Mellman J. 1980. The effect of fuel economy standards on the U.S. automotive market: a hedonic demand analysis. *Transportation Research* **14A**(5–6): 367–378.

Borsch Supan A. 1990. Recent developments in flexible discrete choice models: nested logit analysis versus simulated moments probit analysis. In *Spatial Choices and Processes. Studies in Regional Science and Urban Economics*, Vol. 21, Fischer M *et al.* (eds). North-Holland: Amsterdam; 203–217.

Brownstone D, Bunch D, Golob T, Ren W. 1996. Transactions choice model for forecasting demand for alternative-fueled vehicles. *Research in Transportation Economics* **4**: 87–129.

Brownstone D, Train K. 1999. Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* **28**: forthcoming.

Cardell N, Dunbar F. 1980. Measuring the societal impacts of automobile downsizing. *Transportation Research* **14A**(5–6): 423–434.

Chavas J, Segerson K. 1986. Singularity and autoregressive disturbances in linear logit models. *Journal of Business and Economic Statistics* **4**: 161–169.

Chesher A, Santos-Silva J. 1995. Taste variation in discrete choice models, University of Bristol Working Paper.

Chintagunta P. 1992. Estimating a multinomial probit model of brand choice using the method of simulated moments. *Marketing Science* **11**: 386–407.

Dagsvik J. 1994. Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica* **62**: 1179–1205.

Dubin J, Zeng L. 1991. The heterogeneous logit model, Caltech Social Science Working Paper: 759.

Enberg J, Gottschalk P, Wolf D. 1990. A random effects logit model of work welfare transitions. *Journal of Econometrics* **43**: 63–75.

Feller W. 1966. *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley: New York.

Follman D, Lambert D. 1989. Generalized logistic regression by nonparametric mixing. *Journal of the American Statistical Association* **84**: 295–300.
Formann A. 1992. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* **87**: 476–486.
Gonul F, Srinivasan K. 1993. Modeling multiple sources of heterogeneity in multinomial logit models. *Marketing Science* **12**: 213–229.
Hajivassiliou V, Ruud P. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of Econometrics*, Vol. IV, Engle R, McFadden D (eds); 2384–2441.
Hajivassiliou V, McFadden D. 1997. The method of simulated scores with application to models of external debt crises. *Econometrica* **66**: 273–286.
Heckman J. 1981. Statistical models for discrete panel data. In *Structural analysis of Discrete Data with Econometric Applications*, Manski C, McFadden D (eds). MIT Press: Cambridge; 114–178.
Heckman J. 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process. In *Structural analysis of Discrete Data with Econometric Applications*, Manski C, McFadden D (eds). MIT Press: Cambridge; 179–196.
Heckman J, Singer B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**: 271–320.
Heckman J, Singer B. 1986. Econometric analysis of longitudinal data. In *Handbook of Econometrics*, Vol. III, Griliches Z, Intriligator M (eds). North-Holland: Amsterdam; 1689–1763.
Heckman J, Lochner L, Taber C. 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics* **1**: 1–58.
Jain D, Vilcassim N, Chintagunta P. 1994. A random coefficients logit brand choice model applied to panel data. *Journal of Business and Economic Statistics* **12**: 317–328.
Lee LF, Chesher A. 1986. Specification testing when score test statistics are identically zero. *Journal of Econometrics* **31**: 121–149.
McFadden D, Reid F. 1975. Aggregate travel demand forecasting from disaggregated behavioral models. *Transportation Research Record: Travel Behavior and Values* **534**: 24–37.
McFadden D. 1984. Econometric analysis of qualitative response models. In *Handbook of Econometrics*, Vol. II, Griliches Z, Intriligator M (eds). North Holland: Amsterdam.
McFadden D. 1987. Regression-based specification tests for the multinomial logit model. *Journal of Econometrics* **34**: 63–82.
McFadden D. 1996. Lectures in simulation-assisted statistical inference. University of California, Berkeley, Working Paper.
McFadden D. 1989. A method of simulated moments for estimation of discrete choice models without numerical integration. *Econometrica* **57**: 995–1026.
McFadden D, Ruud P. 1994. Estimation by simulation. *Review of Economics and Statistics* **76**: 591–608.
Montgomery M, Richards T, Braun H. 1986. Child health, breast feeding, and survival in Malaysia: a random effects logit approach. *Journal of the American Statistical Association* **81**: 297–309.
Newey W, McFadden D. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. IV, Engle R, McFadden D (eds). North Holland: Amsterdam; 2111–2245.
Reader S. 1993. Unobserved heterogeneity in dynamic discrete choice models. *Environment and Planning A* **25**: 495–519.
Revelt D, Train K. 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics* **80**: 1–11.
Ruud P. 1996. Approximation and simulation of the multinomial probit model: an analysis of covariance matrix estimation, UC Berkeley Working Paper.
Steckel J, Vanhonacker W. 1988. A heterogeneous conditional logit model of choice. *Journal of Business and Economic Statistics* **6**: 391–398.
Stern S. 1994. Two dynamic discrete choice estimation problems and simulation method solutions. *Review of Economics and Statistics* **76**: 695–702.
Talvitie A. 1972. Comparison of probabilistic modal-choice models. *Highway Research Board Record* **392**: 111–120.
Train K. 1995. Simulation methods for probit and related models based on convenient error patitioning. Working Paper 95-237, Department of Economics, University of California, Berkeley.

Train K. 1998. Recreation demand models with taste differences over people. *Land Economics* **74**: 230–239.
Train K. 1999. Halton sequences for mixed logit. Dept. of Economics, Univ. of California, Berkeley.
Train K, McFadden D, Goett A. 1987. Consumer attitudes and voluntary rate schedules for public utilities. *Review of Economics and Statistics* **69**: 383–391.
Westin R. 1974. Predictions from binary choice models. *Journal of Econometrics* **2**: 1–16.
Westin R, Gillen D. 1978. Parking location and transit demand: a case study of endogenous attributes in disaggregate mode choice models. *Journal of Econometrics* **8**: 75–101.

