A Simple Nonparametric Estimator for the Distribution of Random Coefficients in Discrete Choice Models

Patrick Bajari, Jeremy T. Fox, Kyoo il Kim, and Stephen Ryan*
University of Minnesota and NBER
University of Chicago
University of Minnesota
MIT and NBER

PRELIMINARY VERSION

November 2007

Abstract

We propose an estimator for discrete choice models, such as the logit, with a nonparametric distribution of random coefficients. The estimator, based on linear regression subject to linear inequality constraints, is robust, simple to program, and quick to compute compared to alternative estimators for mixture models. We discuss three methods for proving identification of the distribution of heterogeneity for any given economic model. We prove the identification of the logit mixtures model, which, surprisingly given the wide use of this model over the last 30 years, is a new result. We also derive our estimator's non-standard asymptotic distribution and demonstrate its excellent small sample properties in a Monte Carlo. The estimator we propose can be extended to allow for endogenous prices. The estimator can also be used to reduce the computational burden of nested fixed point methods for complex models like dynamic programming discrete choice.

*Thanks to helpful comments from seminar participants at the AEA meetings, Paris I, Toronto and UCL, as well as Andrew Chesher, Philippe Fevrier, David Margolis and Jean-Marc Robin.

1 Introduction

In this paper, we describe a method for estimating random coefficient discrete choice models that is both flexible and simple to compute. Discrete choice models are concerned with the decision problem of an agent maximizing utility by making a single choice from a set of $j = 1, \ldots, J$ mutually-exclusive options:

$$j = \arg\max \{u_1, \ldots, u_J\},$$

where u_j is the utility of the jth product in a finite, unordered choice set. Indexing each agent by $i = 1, \ldots, N$. Utility is usually specified as a linear function of a $K \times 1$ vector x_j of choice-specific observables and a scalar unobservable shock, $\epsilon_{i,j}$:

$$u_{i,j} = x_{j,t}\beta + \epsilon_{i,j}. \tag{1}$$

In typical industrial organization applications, the econometrician has data on product characteristics and market shares and wants to estimate the vector of the K marginal utilities, β. When N is large in each market, the market share of agents choosing option J is identical to the probability that the utility of choice J is higher than all other alternatives for a single agent

$$s_j = \int 1\left(x_j\beta + \epsilon_{i,j} > x_{j'}\beta + \epsilon_{i,j'}; \forall j' \neq j\right) g\left(\epsilon\right) d\epsilon, \tag{2}$$

where $g\left(\epsilon\right)$ is the joint density of the vector ϵ of the J product and consumer specific unobservables and ϵ is statistically independent of the x_j's. Standard estimators such as maximum likelihood are useful, especially if g is assumed to lie in a parametric family. Following McFadden (1974), in many empirical applications the error terms are assumed to be i.i.d. across consumers and choices with the extreme value distribution $e^{-e^{-\epsilon_j}}$. The functional form assumption gives rise to a simple closed-form solution for the market share:

$$s_j = \frac{\exp\left(x_j\beta\right)}{\sum_{j'=1}^{J} \exp\left(x_{j'}\beta\right)}. \tag{3}$$

Demand elasticities are an important input into studying consumer welfare and the competitiveness of an industry in empirical industrial organization (IO). Suppose that price is characteristic 1 out of K; simple algebra shows that the own-price derivative of market share from the pure logit model is

$$\frac{\partial s_j}{\partial x_{1,k}} = \left(s_j - s_j^2\right)\left(\beta_1\right),$$

which is a function of the market share of product j and not how similar product j is in characteristic space to other products, or how similar the individual components of x_j are to other $x_{j'}$'s. This is a particularly unappealing feature, as a body of prior empirical evidence has shown that products

compete more closely with products that have similar characteristics.

To address this problem, the base logit has been extended to allow for heterogeneity in the marginal utilities across the population of consumers. In these random coefficient models, $u_{i,j} = x_j\beta_i + \epsilon_{i,j}$, where the vector of marginal utilities β_i is a consumer-specific random effect. Market shares are then

$$s_j = \int \frac{\exp\left(x_j\beta\right)}{\sum_{j'=1}^{J}\exp\left(x_{j'}\beta\right)} dF\left(\beta\right), \tag{4}$$

where $F\left(\beta\right)$ is the joint distribution of the marginal utilities. The own- and cross-price elasticities of products in this model are now a function of other products, which generates reasonable substitution patterns across products. See Berry, Levinsohn, and Pakes (1995) for a more in-depth discussion of this point. Again, β is assumed to be statistically independent of ϵ and the J characteristics. Random coefficient logit models have a long history in economics. See Boyd and Mellman (1980), Cardell and Dunbar (1980), Follmann and Lambert (1989), Chintagunta, Jain and Vilcassim (1991), Nevo (2001), Petrin (2002) and Train (2003).

Equation (4) brings up two related issues: computational simplicity in estimation and being flexible about the shape of $F\left(\beta\right)$. On computation, evaluating a likelihood or method of moments objective function based on (4) requires numerical integration of a dimension equal to the number of product characteristics with random coefficients, K. A K-dimensional integral needs to be evaluated $\left(J-1\right)\cdot T$ times, where T is the number of markets with different products. This computational burden of likelihood evaluation then feeds back into the specification of $F\left(\beta\right)$, as researchers adopt parsimonious parametrizations of $F\left(\beta\right)$ in order to reduce the number of parameters used in numerical optimization of the statistical objective function. For example, Berry, Levinsohn and Pakes (1995) specify that each component of β is statistically independent of all other components and that each component has a normal marginal distribution. This computationally-induced parsimony is regrettable, as McFadden and Train (2000) demonstrate that (3) can approximate any system of discrete choice probabilities generated by a random utility model with a linear index $x_j\beta$ structure.

In this paper, we propose a estimator for random coefficient discrete choice models that is both nonparametric with respect to $F\left(\beta\right)$ and computationally simple. The key idea behind the computational simplicity is to choose a parameterization of our model where the dependent variable is a linear function of the structural parameters. We do this by representing heterogeneity in the utility parameters using a large, but finite, number of discrete types. Consider a standard industrial organization setting, where we observe the shares of J products in each of T markets. Suppose that there is a single covariate and tastes are known to lie on the unit interval, and to construct an approximating basis we discretize the preference parameters on the unit interval, $[0,1]$, into R distinct points as follows:

$$\beta^{(r)} = \frac{r}{R} \text{ for } 1 \leq r \leq R.$$

The predicted share of alternative j is then:

$$\bar{s}_{j,t} = \sum_{r=1}^{R} \frac{\exp\left(\beta^{(r)} x_{j,t}\right)}{\sum_{j'=1}^{J} \exp\left(\beta^{(r)} x_{j',t}\right)} \bar{f}\left(\beta^{(r)}\right), \tag{5}$$

where $\bar{f}\left(\beta^{(r)}\right)$ is the mass for type r in the approximation. We put a bar in $\bar{f}\left(\beta^{(r)}\right)$ to emphasize that the mass function is an approximation to $F(\beta)$, see below, and not the density $f(\beta)$. The share equation (5) is itself a discrete type approximation to (4). Let $\widehat{s}_{j,t}$ be the observed share of good j . After adding $\widehat{s}_{j,t}$ to both sides of the above equation and rearranging terms, we obtain:

$$\widehat{s}_{j,t} = \sum_{r=1}^{R} \frac{\exp\left(\beta^{(r)} x_{j,t}\right)}{\sum_{j'=1}^{J} \exp\left(\beta^{(r)} x_{j',t}\right)} \bar{f}\left(\beta^{(r)}\right) + (\widehat{s}_{j,t} - \bar{s}_{j,t}). \tag{6}$$

The above equation suggests that we can estimate $\bar{f}\left(\beta^{(r)}\right)$ by linear regression. Keep in mind that $\beta^{(r)}$ and hence $\frac{\exp\left(\beta^{(r)} x_{j,t}\right)}{\sum_{j'=1}^{J} \exp\left(\beta^{(r)} x_{j',t}\right)}$ are fixed ahead of time by the choice of the approximating basis. The dependent variable in our regression is the observed share $\widehat{s}_{j,t}$, the regressors are the predicted shares under each of the R types, $\frac{\exp\left(\beta^{(r)} x_{j,t}\right)}{\sum_{j'=1}^{J} \exp\left(\beta^{(r)} x_{j,t'}\right)}$, and the error term is $(\widehat{s}_{j,t} - \bar{s}_{j,t})$. Note that the error term is pure additive measurement error. To some, this estimator may seem obvious, but we feel there is great value in exploring its properties fully in this paper.

Typically, we get a better approximation to the true distribution of types F if we impose that the weights form a probability mass function: $\bar{f}\left(\beta^{(r)}\right) \geq 0 \, \forall \, r$ and $\sum_{r=1}^{R} \bar{f}\left(\beta^{(r)}\right) = 1$. Our estimator is just linear least squares with linear inequality constraints, a well known numerical problem with available, custom routines in packages such as Matlab (lsqlin). We can then form an approximation to the true distribution of types using the estimated distribution of types

$$\hat{F}(\beta) = \sum_{r=1}^{R} 1\left[\beta^{(r)} \leq \beta\right] \bar{f}\left(\beta^{(r)}\right). \tag{7}$$

Even if the K individual components of the basis vectors $\beta^{(r)}$ for a given r are chosen to be statistically independent, the estimated distribution $\hat{F}(\beta)$ is unlikely to satisfy statistical independence across its K arguments. Thus, $\hat{F}(\beta)$ is a multidimensional, nonparametric distribution estimator.

The linear regression estimator has three important properties. First, it is simple to compute even if we include a very large number (e.g. a thousand) of basis points. In contrast to the standard approaches in the discrete choice literature our estimator does not require a nesting a high-dimensional integration inside the computation of our statistical objective function. Second, inequality constrained linear least squares is a convex optimization problem and so appropriate routines will always find a global minimum. Third, if we include a very large number of basis points to discretize the unit interval, we will have a very flexible model of the distribution of β. Our approach to approximating the unknown distribution $F(\beta)$ may be less efficient than estimators

based on alternative approximation schemes, but we think this possible theoretical disadvantage is more than outweighed by the computational simplicity and generality of our approach. A Monte Carlo analysis demonstrates that our method is able to estimate quite complicated distributions of random coefficients with a minimal computational burden. In a related paper, Bajari, Fox, and Ryan (2007), we provide a Monte Carlo example where the loss of statistical efficiency relative to a parametric maximum likelihood estimator, when the parametric assumptions are correct, is inconsequential in a typical discrete choice framework. Also, the Monte Carlos show the MLE makes false predictions when its parametric assumptions are violated.

In practice, approximating the distribution of types using a discrete grid of points with equal spacing may not provide the best approximation. We discuss some alternative ways to approximate the distribution of β using mixtures of normal distributions and location scale models. Using normal mixtures will generate a smooth estimated density for β. Location scale models address the concern that the support of the random coefficients may not be known before estimation. We show how to modify our estimator for data on choices by individual consumers.

A common problem in empirical IO applications is that a key product characteristic, price, is correlated with an aggregate demand shock, often called the unobserved product characteristic ξ_j. Berry, Levinsohn and Pakes (1995) show how to use instrumental variables in a nonlinear discrete choice model. As endogeneity is important in applications, we discuss replacing the Berry et al. parametric independent normal $F(\beta)$ with our nonparametric specification.

Our idea is more general than static, single agent discrete models. The key idea of estimating distributions of heterogeneity using a linear regression can be applied to many other economic models. We briefly discuss dynamic programming discrete choice models. Adding random coefficients to these models in applications such as demand for storable goods is usually portrayed as computationally burdensome (Erdem, Imai and Keane 2003, Hendel and Nevo 2006). Similarly to the parametric estimator of Ackerberg (2001), our estimator has the computational advantage that the economic model (a dynamic program) can be solved for a finite number of parameter values (the basis vectors) before numerical optimization begins. This can be a tremendous computational savings over the standard Rust (1987) approach of solving the model at every call to a statistical objective function by an optimization routine.

A key question that has not been adequately addressed in the literature is identification: under what conditions is there a unique $F(\beta)$ that solves (4)? If $F(\beta)$ is not identified, there is little purpose in proposing an estimator as any estimator would be inconsistent. We discuss identification and provide conditions for the identification of $F(\beta)$ in economic and statistical models. We introduce three non-primitive conditions that ensure identification of continuous mixtures like $F(\beta)$. The first is linear independence, the second is bounded completeness of the density function, and the third is a condition that says a density is uniquely determined by its moments. The third condition is particularly novel and useful for showing identification of the mixture distributions

for many economic models, not just discrete choice and random coefficients logit. In fact, the identification theorems and the asymptotic results to follow use the general notation

$$P(x) = \int g(x, \beta)\, dF(\beta)$$

where $g(x, \beta)$ is a model indexed by covariates x and parameters β, $F(\beta)$ is the mixing distribution, and $P(x)$ is the probability of some outcome. This notation encompasses the random coefficients logit and many more economic and statistical models. We do, however, specifically prove the the identification of the logit mixtures model, which, surprisingly given the wide use of this model over the last 30 years, has never been proved. Our identification result requires that all product characteristics be continuous. The characteristics only need to vary locally; we do not use identification at infinity.

Another technical contribution is to derive the asymptotic distribution of our estimator. The distribution we derive is nonstandard because of a boundary problem: some weights can be zero in the true data generating process. We derive asymptotic distributions for the case where there is only a finite number of types. We derive rates of convergence for the case where there is a possibly uncountable number of types. We include several more advanced results, such as optimally combining different estimators, estimating jointly with parameters that are not random coefficients, and estimating adaptively when many parameters may have a true value of zero.

Our estimator is a mixtures estimator that could be applied to many models. There are many parametric frequentist and Bayesian mixtures estimators in the literature. The most common frequentist, nonparametric estimator is nonparametric maximum likelihood or NPMLE (Laird 1978, Böhning 1982, Lindsay 1983, Heckman and Singer 1984). The NPMLE likelihood is hard to numerically optimize; often the EM algorithm (as opposed to standard solvers available in commercial packages) is required and even the EM algorithm is not guaranteed to find the global maximum. Indeed, the literature has worried about the strong dependence of the output of the EM algorithm on initial starting values, much more so than for other methods (Verbeek, Vlassis and Kröse 2002). Because of these computational issues, it is not common to use NPMLE to estimate random coefficient discrete choice models, despite the estimator's long history. Indeed, Li and Barron (2000) introduce a likelihood mixtures iterative estimator that is easier to compute than full NPMLE and has many of the same attractive statistical properties. This iterative procedure, however, involves many nonlinear optimizations of potentially non-concave likelihoods; it is difficult to imagine the computational cost of trying many starting values at each iteration of a potentially long process. It is certainly much more costly to program than our estimator. Finally, the Li and Barron iterative procedure will not be computationally simple at all when combined with some tools used by economists: instrumental variables like Berry, Levinsohn and Pakes (1995) and dynamic programming like Rust (1987). Our method retains some of its computational simplicity for these complex structural models.

More recent classical methods in the statistics literature include Barbe (1998) and Roueff and Ryden (2005). These estimators are not as computationally simple as ours, and so are less likely to be used by economists. Rossi, Allenby and McCulloch (2006) introduce a very flexible (for smooth distributions) but computationally-intensive Bayesian estimator for mixtures distributions, particularly the random coefficients logit. While innovative, their MCMC simulation methods require lots of user training and also lots of computer time to simulate the ergodic distribution of the parameters in an approximating mixture of normals. Also, the particular economic model being mixed will have to inform the method for specifying priors and random number generation, details that require new math to use MCMC methods for new economic models (Zeithammer and Lenk 2006).

Compared to existing nonparametric mixtures estimators, our estimator's main advantage is computational simplicity, not statistical efficiency. Rather than jointly estimating the weights on mixture components and the components themselves, we specify a very large number (R) of components and estimate only the weights (in the simplest form of the estimator; we propose variants). Computation is evidently a binding constraint in practice; these methods are rarely used for complex economic models, in IO or other fields. Our estimator is also simpler than alternative schemes for approximating $f(\beta)$, such as families of orthogonal polynomials and neural networks.

The discussion of nonparametric alternatives should not obscure the Monte Carlo evidence in Bajari, Fox, and Ryan (2007) that shows that using our estimator results in only a minimal loss of efficiency compared to a parametric MLE model when the parametric assumptions are correct. Together, this suggests that any statistical efficiency gain of NPMLE over our estimator must be small, as NPMLE's performance must fit between the efficient parametric estimator's performance and ours. Also, we showed the parametric model can produce very biased estimates when its assumptions are violated. Using most any nonparametric estimator is likely preferable to using a parametric estimator.

We are interested in the structural interpretation of the mixtures density $f(\beta)$ as the distribution of preference parameters among consumers. This distribution can be used to compute the consumer surplus for introducing a new good, for example (Petrin 2002). There are many other schemes that could provide a flexible reduced-form approximation to a market shares equation; we explicitly wish to rely on fully structurally estimating the parameters of a well-specific economic model.

There is a smaller, non-mixtures literature on estimation of discrete choice models with random coefficients. For the special case of a binary ($J = 2$) discrete choice model (not a general mixtures model), Ichimura and Thompson (1998) introduce a maximum likelihood estimator for the distribution of random coefficients (without relying on the logit model) that is more difficult to implement than our linear regression estimator. Their method only allows continuous covariates in $x_{j,t}$. Lewbel (2000) considers the identification and estimation of $E[\beta]$ for $J \geq 2$ using a large support (identification at infinity) special regressor and a mean independence assumption on the other

$K-1$ product characteristics. His specification nests random coefficients on the $K-1$ non-special regressors, but his multinomial-choice estimator requires high-dimensional nonparametric density estimates involving the product characteristics. Manski's (1975) maximum score estimator is consistent for $E[\beta]$ in the presence of random coefficients for the case of $J=2$ choices only. Briesch, Chintagunta and Matzkin (2007) replace the linear index $x'\beta$ with a nonparametric function of x, but allow a random coefficient on only an intercept term.

The paper is organized as follows. In Section 2 we discuss the variants of the estimator in detail. We discuss identification in Section 3 and provide asymptotic theory for our estimator in Sections 4 and 5. We demonstrate the small-sample performance of our estimator in a Monte Carlo in Section 6. Section 7 concludes.

2 The Estimator

2.1 Linear Regression

Say we have data on $t=1\dots T$ markets. Each market has J products, and the characteristics of products vary across markets. Equation (6) suggests that it is possible to compute the approximation weights using regression. That is[1]:

$$\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R} = \arg\min_{\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}} \sum_{t=1}^{T}\sum_{j=1}^{J-1}\left(\hat{s}_{j,t} - \sum_{r=1}^{R}\frac{\exp\left(x_{j,t}'\beta^{(r)}\right)}{\sum_{j'=1}^{J}\exp\left(x_{j',t}'\beta^{(r)}\right)}\bar{f}(\beta^{(r)})\right)^2. \tag{8}$$

The values

$$\frac{\exp\left(x_{j,t}'\beta^{(r)}\right)}{\sum_{j'=1}^{J}\exp\left(x_{j',t}'\beta^{(r)}\right)}$$

are known and fixed by the choice of basis vectors. This is the ordinary least squared objective function, with closed form $(Z'Z)^{-1}(Z'Y)$, where Y is a $JT\times 1$ vector of market shares and Z is a $JT\times R$ matrix of predicted market shares, where the r-th column of Z corresponds to the predicted shares in each market under parameter vector $\beta^{(r)}$.[2] The closed form solution exists when $Z'Z$ is nonsingular, which is only possible if the number of basis vectors is less than or equal to the number of observations on market shares. If there is not a unique solution, then there are just two or more equally valid (by the least squares fit criterion) constructions of $\hat{F}(\beta)$, as seen in (7), given the data and the choice of basis vectors. Also, there is no need to include a constant term in the regression.

In discrete approximations, researchers sometimes allow as few as two or three types. As our types are picked beforehand and the approximation weights enter linearly, we can easily allow R

[1] Note that we have at most $J-1$ independent shares since $\sum_{j=1}^{J}(\widehat{s}_{j,t}-\bar{s}_{j,t})=1-1=0$.

[2] Finding the efficient weighting matrix for our estimator involves constructing a feasible generalized least squares estimator.

to be in the hundreds or thousands. The computational cost of increasing R is the same as adding regressors to a linear regression.

Note $\hat{s}_{j,t}$ is the measured share in $t = 1, ..., T$. The "error term" in our regression is the pure measurement error $(\widehat{s}_{j,t} - s_{j,t})$. In all markets there are only a finite number of individuals. Therefore, $\hat{s}_{j,t}$ is a sample mean with sampling error in it. Central limit theorem arguments suggest that the estimation error in a sample mean will be uncorrelated with the true mean, across observations on markets and products. In notation, $(\widehat{s}_{j,t} - s_{j,t})$ and $\frac{\exp(x'_{j,t}\beta^{(r)})}{\sum_{j'=1}^{J}\exp(x'_{j',t}\beta^{(r)})}$ have zero covariance and the condition for the consistency (in the number of markets T) of least squares is satisfied.[3]

2.2 Constrained Least Squares

In finite samples, individual estimated regression coefficients $\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}$ may not be greater than zero and may not sum to one exactly. As a result, the approximation (7) is not a well-defined distribution. This makes it hard to use $\hat{F}(\beta)$ structurally to, say, compute moments of the estimated random coefficient distribution. In terms of prediction, a model using $\hat{F}(\beta)$ may predict market shares which do not sum to one for certain values of x.

One may enforce the constraints that

$$\begin{aligned} \sum_{r=1}^{R} \tilde{f}\left(\beta^{(r)}\right) &= 1 \\ \tilde{f}\left(\beta^{(r)}\right) &\geq 0 \text{ for all } r. \end{aligned}$$

In this case, our estimator minimizes (8) subject to these constraints. This minimization problem is a quadratic programming problem subject to linear inequality constraints.[4] The minimization problem is convex and routines like Matlab's lsqlin guarantee finding a global optimum. Again, if multiple global optimums exist, they all form equally valid approximations $\hat{F}(\beta)$.

One way to remove the R inequality constraints $\tilde{f}\left(\beta^{(r)}\right) \geq 0$ is to maximize with respect to the change of variables $w^{(r)} = \log\left(\tilde{f}\left(\beta^{(r)}\right)\right)$, which can be both positive or negative when $\tilde{f}\left(\beta^{(r)}\right) \geq 0$. The transformed variables $w^{(r)} = \log\left(\tilde{f}\left(\beta^{(r)}\right)\right)$ enter the least squares objective function nonlinearly as $\tilde{f}\left(\beta^{(r)}\right) = \exp\left(w^{(r)}\right)$, so nonlinear least squares should be used. The nonlinear equality constraint $\sum_{r=1}^{R}\exp\left(w^{(r)}\right) = 1$ can optionally be imposed. The extra nonlinearities often make finding the global minimum of the least squares objective function harder than linear least squares, but without the possibility than an inequality constraint binds the central limit theorem (at least for $R-1$ of the weights) applies because parameters will not be on the boundary

[3] When we have individual level data, the $\widehat{s}_{j,t}$ will be a set of indicator variables that are equal to 1 if individual t chooses option j and 0 otherwise.

[4] Statistical inference for linear regression subject to a set of inequality constraints has been studied by Judge and Takayama (1966), Liew (1976), Geweke (1986), and Wolak (1987).

of the parameter space, as we discuss later. We suspect most users will prefer the linear specification.

Some researchers may want to impose more constraints on the density $F(\beta)$ itself. One could impose, using additional constraints, that two components of the random vector β have zero covariance, for example, by deriving the formula for the covariance using the discrete-type approximation, $\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}$. These economic constraints would likely be nonlinear and increase the difficulty of computation.

2.3 Smooth Basis Densities

If the true distribution of random coefficients is smooth, it may be desirable to use smooth basis functions rather than the discrete points described above. A mixture of normal distributions can approximate an arbitrary smooth density function with full support. We let our basis points be indexed by $\left\{\mu^{(r)}, \sigma^{(r)}\right\}_{r=1}^{R}$ where $\mu^{(r)}$ is a $K \times 1$ vector of mean parameters for the K product characteristics and $\sigma^{(r)}$ is a $K \times 1$ vector of standard deviations. Let $N\left(\beta_k|\mu_k^{(r)}, \sigma_k^{(r)}\right)$ denote the density of the k-th random coefficient using the basis points $\mu_k^{(r)}, \sigma_k^{(r)}$. We impose independence in the basis functions, but the estimated final density $\hat{f}(\beta)$ will not be independent.

Under normal basis functions, the joint density for a given r is just the product of the marginals, or

$$N\left(\beta|\mu^{(r)}, \sigma^{(r)}\right) = \prod_k N\left(\beta_k|\mu_k^{(r)}, \sigma_k^{(r)}\right). \tag{9}$$

Let $\bar{f}(r)$ denote the probability mass associated with the basis density parameterized by $\left\{\mu^{(r)}, \sigma^{(r)}\right\}$, in the approximation. As in the previous section, we enforce the constraint that $\left\{\bar{f}(r)\right\}_{r=1}^{R}$ generates a well defined mass function. That is,

$$\sum_{r=1}^{R} \bar{f}(r) = 1 \tag{10}$$

$$\bar{f}(r) \geq 0 \text{ for all } r. \tag{11}$$

The estimated density of random coefficients is then:

$$\hat{f}(\beta) = \sum_{r=1}^{R} \bar{f}(r) N\left(\beta|\mu^{(r)}, \sigma^{(r)}\right) \tag{12}$$

$$= \sum_{r=1}^{R} \bar{f}(r) \prod_k N\left(\beta_k|\mu_k^{(r)}, \sigma_k^{(r)}\right). \tag{13}$$

From this density, integration gives a smooth distribution function.

The mixture of normal distributions results in a smooth estimate of the underlying density of marginal utilities, while the discrete models used in the previous subsections do not have continuous

density functions. Using a smooth estimator is likely to improve the efficiency if the true underlying density is smooth. Using a density certainly improves the visualization of the distribution of heterogeneity, as many humans have a hard time interpreting three-dimensional plots of two-variable CDFs.

To estimate the weights, we need the logit probabilities for continuum of types with the basis density $N\left(\beta|\mu^{(r)},\sigma^{(r)}\right)$, or

$$P_t^{(r)} = \int \dots \int \frac{\exp\left(x'_{j,t}\beta\right)}{\sum_{j'=1}^{J}\exp\left(x'_{j',t}\beta\right)} \left(\prod_{k=1}^{K} N\left(\beta_k|\mu_k^{(r)},\sigma_k^{(r)}\right)\right) d\beta_1 \dots d\beta_K. \tag{14}$$

This $K-$dimensional numerical integral is solved only once, as the $\left\{\mu_k^{(r)},\sigma_k^{(r)}\right\}_{k=1}^{K}$ for each basis element r are chosen before numerical optimization begins. The minimization problem to estimate the weights on the basis approximation is

$$\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R} = \arg \min_{\left\{\bar{f}:\bar{f}\left(\beta^{(r)}\right)\geq 0\,\forall r,\sum_{r=1}^{R}\tilde{f}\left(\beta^{(r)}\right)=1\right\}} \sum_{t=1}^{T}\sum_{j=1}^{J-1}\left(\widehat{s}_{j,t} - \sum_{r=1}^{R} P_t^{(r)}\bar{f}\left(r\right)\right)^2.$$

This is a linearly constrained linear regression using a data matrix with elements $P_t^{(r)}$. Adding smooth basis densities does not change the difficulty of the constrained least squares problem as $P_t^{(r)}$ is computed in a first stage.

2.4 Location Scale Model

In many applications, the econometrician may not have good prior knowledge about the support region where most of the random coefficients lie. The marginal utilities in a logit model are expressed using an arbitrary scale normalization: the standard deviation of the ϵ_{ijt} taste shocks is $\pi/\sqrt{6} \approx 1.3$. Given the arbitrary units, researchers might have little idea about how to pick the support of the basis vectors ahead of time.[5] It may be useful to search over a set of location and scale parameters that determine the general neighborhood where our basis points lie.

Return to the discrete basis vectors estimator. Let the unscaled basis vectors $\left\{\beta^{(r)}\right\}_{r=1}^{R}$ lie in the set $[0,1]^K$, that is, the K-fold Cartesian product of the unit interval. We include a set of location and scale parameters μ_k and σ_k, $k = 1, ..., K$ and define the rth random coefficient for the kth characteristic as $\mu_k + \sigma_k\beta_k^{(r)}$.

In numerical optimization, we now search over $R+2K$ parameters corresponding to $\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}$,

[5] One suggestion is to estimate a logit model without random coefficients first and use the estimates of the marginal utilities as the means of the support. Unfortunately, the homogeneous logit is an estimator with sampling error in it, and this two-step procedure requires adjusting the standard errors of the second-stage for pre-testing.

$\mu = (\mu_1, ..., \mu_K)'$ and $\sigma = (\sigma_1, ..., \sigma_K)'$. Market shares predictions for type r are:

$$\frac{\exp\left(\sum_{k=1}^{K} x_{k,j,t}\left(\mu_k + \sigma_k \beta_k^{(r)}\right)\right)}{\sum_{j'=1}^{J} \exp\left(\sum_{k=1}^{K} x_{k,j,t'}\left(\mu_k + \sigma_k \beta_k^{(r)}\right)\right)}. \tag{15}$$

This cannot be precomputed, as the location and scale parameters need to be estimated. Our estimator for the weights solves the following nonlinear least squares problem

$$\min_{\mu,\sigma,\left\{\bar{f}:\bar{f}\left(\beta^{(r)}\right)\geq 0\,\forall r,\sum_{r=1}^{R}\left(\beta^{(r)}\right)=1\right\}} \sum_{t=1}^{T}\sum_{j=1}^{J-1}\left(\widehat{s}_{j,t} - \sum_{r=1}^{R} \frac{\exp\left(\sum_{k=1}^{K} x_{k,j,t}\left(\mu_k + \sigma_k \beta_k^{(r)}\right)\right)}{\sum_{j'=1}^{J} \exp\left(\sum_{k=1}^{K} x_{k,j,t'}\left(\mu_k + \sigma_k \beta_k^{(r)}\right)\right)} \bar{f}\left(\beta^r\right)\right)^2. \tag{16}$$

The appropriate Matlab routine is lsqnonlin, and there is no theorem that the routine will converge to a global minimum.[6] Our numerical experience is that solving this minimization problem using a canned routine is much easier than the custom programming needed to apply the EM algorithm to the NPMLE.

The location scale approach can be combined with smooth densities for the basis functions. When the location and scale parameters are not fixed before optimization, $P_t^{(r)}$ defined in (14) cannot be precomputed. As $P_t^{(r)}$ involves a numerical integral, combining the location scale method with smooth densities presents a computational downside. As part of non-linear optimization, $R \cdot J \cdot T$ numerical integrals of dimension K will need to be solved each time the least squares objective function is evaluated, although the linear constraints remain the same.

2.5 Product Characteristic Endogeneity

In applied work, the vector $x_{j,t}$ may not contain all of the characteristics of product j that are observed by the consumer. Let the scalar $\xi_{j,t}$ denote such an omitted product attribute or aggregate demand shock. The total utility of a choice is now $u_{i,j,t} = x'_{j,t}\beta + \xi_{j,t} + \epsilon_{i,j,t}$. One role of an aggregate error term such as $\xi_{j,t}$ is to give positive support to the data. With many consumers in each market so there is no sampling in shares, $\xi_{j,t}$ will give positive support on all vectors of J market shares that sum to 1. Using individual data does not preclude the need to include $\xi_{j,t}$, as such aggregate shocks are likely present in many applications.

Our previous estimator will be inconsistent if $\xi_{j,t}$ enters the model at all, but especially if $\xi_{j,t}$ is correlated with one or more of the observed product characteristics. This is likely as price and quantity are determined simultaneously in market equilibrium. Recent applied work has found

[6] Minimization may be faster using an interface to nonlinear solvers, such as AMPL, that supports automatic differentiation. Also, μ and σ do not even enter the constraints, which are still linear.

that the bias from a correlation of the observed price with an unobserved product characteristic may be quite severe (see Berry, Levinsohn and Pakes (1995), Nevo (2001) and Petrin (2002)). As the coefficient on price is the key parameter governing both demand elasticities and measures of consumer welfare (say from the introduction of a new good), correcting this bias is essential.

Instruments $z_{j,t}$ can be used to resolve this omitted variable bias problem. Consistency requires a mean independence assumption: $E\left[\xi_{j,t} \mid z_{j,t}\right] = 0$. A GMM objective function, the empirical analog to $E\left[\xi_{j,t} \mid z_{j,t}\right]$, forms the basis for estimation. As $\xi_{j,t}$ enters the model nonlinearly, a numerical method must be used to compute the implied $\hat{\xi}_{j,t}$, which is a function of guesses for the model's structural parameters as well as the market share and product characteristic data.

Berry, Levinsohn and Pakes (1995), commonly known as BLP, introduce one such numerical method and illustrate it with a parametric normal distribution for $F(\beta)$. However, the basic approach extends to our finite-type approximation to a nonparametric $F(\beta)$. With L instruments in each $z_{j,t}$, a researcher minimizes the (here unweighted) GMM objective function

$$\min_{\left\{\bar{f}:\bar{f}\left(\beta^{(r)}\right)\geq 0,\sum_{r=1}^{R}\bar{f}\left(\beta^{(r)}\right)=1\right\}} \sum_{l=1}^{L}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J-1} z_{l,j,t}\xi_{j,t}\left(\hat{F}\left(\beta\right)\right)\right)^{2},$$

where $\hat{F}(\beta)$ is based on $\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}$ from (7) and $\xi_{j,t}\left(F(\beta)\right)$ is a function that finds the $\xi_{j,t}$ implied by an arbitrary random coefficient distribution and the market share data. In other words, $\{\xi_{j,t}\}_{j=1}^{J}$ for market t solves the J nonlinear market share equations

$$s_{j,t} = \sum_{r=1}^{R}\frac{\exp\left(\beta^{(r)}x_j + \xi_{j,t}\right)}{\sum_{j'=1}^{J}\exp\left(\beta^{(r)}x_{j'} + \xi_{j',t}\right)}\bar{f}\left(\beta^{(r)}\right). \tag{17}$$

BLP prove that a unique solution exists (their Appendix 1 Theorem). For BLP's strategy to work without much bias in a finite sample, $s_{j,t}$ must have no measurement or sampling error in it, as $\xi_{j,t}$ must be exactly invertible from this system of equations.[7] Evaluating the GMM objective function is time consuming because the contraction mapping BLP use to solve the system converges at only a linear rate.[8] Fox and Su (2007) argue that researchers who set the tolerance of the contraction mapping low to speed its completion introduce numerical error that can lead to false estimates of economic importance. Fox and Su argue that jointly minimizing the GMM objective function over $\left\{\bar{f}\left(\beta^{(r)}\right)\right\}_{r=1}^{R}$ and $\left\{\{\xi_{j,t}\}_{j=1}^{J}\right\}_{t=1}^{T}$ while enforcing (17) as nonlinear constraints in the optimization has better numerical properties and is much faster than computing $\left\{\{\xi_{j,t}\}_{j=1}^{J}\right\}_{t=1}^{T}$ at each call to

[7] Berry, Linton and Pakes (2006) present Monte Carlo evidence that measurement and sampling error in $s_{j,t}$ introduces a large, finite sample bias in the estimates of moments of $F(\beta)$.

[8] In their parametric normal model for $F(\beta)$, BLP actually invert $\delta_j = E\left[\beta x_j\right] + \xi_{j,t}$ rather than $\xi_{j,t}$. With their normality assumption, this allows them to solve for $E\left[\beta\right]$, the location parameters of the normal distribution, inside the GMM objective function. This reduces the number of parameters being minimized by the optimization routine.

the objective function.[9]

2.6 Dynamic Programming Models

The random coefficient logit is not the only model where heterogeneity could be important. The methods that we have described above can be applied to more general discrete choice models including dynamic programming discrete choice models, as in Rust (1987, 1994). Adding random coefficients to a dynamic discrete choice problem appears to be a computational obstacle in some of the applied literature, say on consumer inventory decisions and responses to temporary price changes (Erdem, Imai and Keane 2003, Hendel and Nevo 2006).[10] Here we show that our nonparametric estimator is still computationally simple.

Consider adding random coefficients to the homogeneous-coefficient dynamic discrete choice models discussed in Rust. In this framework, agents maximize expected discounted profits by solving a dynamic programming model. The flow utility of agent i in a period t from choosing action j is:

$$u_{ij} = x'_{j,t}\beta_i + \varepsilon_{ijt}$$

The error term ε_{ijt} is a preference shock for agent i's utility to choice j at time period t. The error term is iid extreme value across agents, choices and time periods. For simplicity, make ϵ_{ijt} logit. Agent i's decision problem is dynamic because there is a link between the current and future values of x_t through current decisions. Let $\pi_\theta\left(x_{t+1} \mid x_t, j\right)$ denote the transition probability for the state variable x as the function of a finite number of parameters θ. This does not involve random coefficients and so θ can usually be estimated in a preliminary stage.

The goal is to estimate $F\left(\beta\right)$, the distribution of the random coefficients. Again we draw R basis vectors $\beta^{(r)}$. For each of the R basis vectors, we can solve the corresponding single-agent dynamic programming problem for the state-x_t value functions, $V^{(r)}\left(x_t\right)$. Solving the dynamic programming problem for one type r is equivalent to solving a system of nonlinear equations with as many equations as states. Once all value functions $V^{(r)}\left(x_t\right)$ are known, the choice probabilities $P^{(r)}\left(j \mid x, \beta^{(r)}\right)$ for all combinations of choices j and states x can be calculated as

$$P^{(r)}\left(j \mid x, \beta^{(r)}\right) = \frac{\exp\left(x'_{j,t}\beta^{(r)} + \delta E\left[V^{(r)}\left(x_{t+1}\right) \mid x_t, j\right]\right)}{\sum_{j'=1}^{J}\exp\left(x'_{j',t}\beta^{(r)} + \delta E\left[V^{(r)}\left(x_{t+1}\right) \mid x_t, j'\right]\right)} = \frac{\exp\left(v^{(r)}\left(j, x_t\right)\right)}{\sum_{j'=1}^{J}\exp\left(v^{(r)}\left(j', x_t\right)\right)},$$

[9] Let the only endogenous product characteristic be price, $p_{j,t}$. Petrin and Train (2003) make an additional assumption about the price reduced form. They assume that $p_{j,t} = \mathbf{p}\left(z_{j,t}, \theta\right) + \xi_{jt}$ and $E\left[\xi_{jt} \mid z_{j,t}\right] = 0$. Then non-linear least squares can be used to estimate θ and the residual from this regression is an estimate of $\xi_{j,t}$. This residual $\hat{\xi}_{j,t}$ can then be substituted into the demand model and our estimator in Section 2.1 can be used, treating $\hat{\xi}_{j,t}$ as an observable product characteristic. Using an estimate $\hat{\xi}_{j,t}$ rather than the true $\xi_{j,t}$ has the potential to introduce finite-sample bias that only goes away as $\hat{\theta} \to \theta$ in the price reduced form.

[10] These papers also deal with unobserved, time persistent consumer state variables (inventories). Our methods could be combined with theirs to address unobserved state variables as well, but the details are a little too much for this paper.

where $v^{(r)}(j, x_t)$ is Rust's choice-specific continuation value, here implicitly defined in the above equation. Only R dynamic programming problems need to be solved (unless the location scale or smooth densities methods are being used).

If we observe market shares (estimated choice probabilities) of each j and x, we then perform the minimization in (8), with or without constraints. The elements of the data matrix for each type are $P^{(r)}\left(j \mid x, \beta^{(r)}\right)$, the choice probabilities from solving the model in the first stage. This method is more computationally feasible than Rust (1987, 1994), whose likelihood estimator requires that a dynamic program be solved once for each call to the likelihood function by the optimization routine. We require only R dynamic programming solutions and are nonparametric on the distribution of random coefficients. Ackerberg (2001) also requires only R dynamic programming solutions and allows random coefficients, but is parametric on the distribution $F(\beta)$.

2.7 Individual Data

Any nonparametric estimator, including ours, requires lots of data. Some IO datasets, such as data on the US car market in Berry, Levinsohn and Pakes (1995), have hundreds of products with detailed characteristics. Therefore, our estimator may be appropriate for these settings. Other datasets may have geographically or temporally distinct markets.

However, there are also datasets listing consumer purchases and consumer demographics. By interacting consumer demographics with product characteristics to construct covariates $x_{i,j,t}$, a researcher can observe variation in the x's that does not require data on different markets. In fake data experiments, we have found our estimator produces good estimates for individual data when we use a linear probability model. The dependent variable $y_{i,j,t}$ for consumer i, product j and market t is

$$y_{i,j,t} = \begin{cases} 1 & \text{if } i \text{ picks } j \\ 0 & \text{otherwise} \end{cases}.$$

One then uses the linear probability model that regresses of $y_{i,j,t}$ on the logit choice probabilities

$$\frac{\exp\left(x_{i,j,t}\beta^{(r)}\right)}{\sum_{j'=1}^{J}\exp\left(x_{i,j't}\beta^{(r)}\right)}$$

for the R types evaluated at consumer i's demographics and the J product characteristics for market t.[11] With T markets, J products per market, and N consumers per market, the total number of rows in our linear regression data matrix is $J \cdot T \cdot N$. Bajari, Fox and Ryan (2007) mention some extensions of the linear probability model approach to individual panel data.

[11] In our simplest setup, we assume consumer demographics are independent of random coefficients. However, there can be random coefficients on the interaction between consumer demographics and product characteristics in $x_{i,j,t}$. Alternatively, we could condition all aspects of the model on observable consumer characteristics and estimate a mixtures density that is parameterized by consumer demographics w_i: $F(\beta \mid w)$.

The linear probability model applied to our multinomial choice model with individual data works well but is somewhat ad hoc from a statistical theory standpoint, as it is not clear that there is an error term justifying the linear probability specification. Alternatively, one could use our mixtures specification in a log-likelihood

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{J} y_{i,j,t}\log\left(\sum_{r=1}^{R}\frac{\exp\left(x_{i,j,t}\beta^{(r)}\right)}{\sum_{j'=1}^{J}\exp\left(x_{i,j't}\beta^{(r)}\right)}\bar{f}\left(\beta^{(r)}\right)\right).$$

Computationally, we can still pick the R types $\beta^{(r)}$ on a grid and estimate only the weights $\bar{f}\left(\beta^{(r)}\right)$. The equality constraint $\sum_{r=1}^{R}\bar{f}\left(\beta^{(r)}\right)=1$ must be imposed in optimization to avoid an infinite likelihood. This likelihood estimator is a rigorous multinomial choice model estimator. The weights $\bar{f}\left(\beta^{(r)}\right)$ enter the log-likelihood nonlinearly, but this likelihood will still be more regular and so much easier to numerically maximize than if the types $\beta^{(r)}$ and weights $\bar{f}\left(\beta^{(r)}\right)$ were jointly estimated, as in the NPMLE discussed in the introduction.

2.8 Choosing R

We have discussed the location and scale model for when the support of $\beta^{(r)}$ is not known and the $\beta^{(r)}$ are, say, picked on a grid. It is also natural to ask how to choose R, the number of basis components. As the nonparametric rates of convergence in Section 5 are asymptotic, they are less useful for choosing R in a finite sample. However, the rates do suggest that that R should not increase too quickly with T, the number of observations. In fake data experiments, we have found a small R (say a few hundred basis vectors) approximation to $F(\beta)$ can have good finite-sample performance.

From the viewpoint of statistical inference, we must caution against an iterative procedure where one chooses R grid points, discards those with zero points, and chooses new grid points near the apparent areas of the support of β with more mass. As statistical sampling error enters the first set of estimates, this iterative procedure introduces pre-testing and will alter the asymptotic distribution.

3 Identification

In this section, we consider the problem of identifying the density of random coefficients $f(\beta)$. We shall consider a more general model which includes the random coefficients logit as a special case, as well as many models that are not discrete choice models. We assume that $g(x,\beta,\varepsilon)$ is a known function which maps covariates, x, random coefficients, β, and unobservable shocks, ϵ, into outcomes. We do not focus on the unobservable shocks ϵ in the logit model: we work with $g(x,\beta)=\int g(x,\beta,\epsilon)\,dH(\epsilon)$, where $H(\epsilon)$ is the known distribution of shocks. Alternatively, the shocks could be placed into β and their identification explored.

The econometrician observes covariates x and the probability of some binary outcome, $P(x)$. For a model with a more complex outcome (including a continuous outcome y), we can always consider whether some event ($y < \frac{1}{2}$ say) happened or did not happen. $P(x)$ is the probability of the event happening. We can fit the mixed logit models of the previous section into this framework by defining the terms as

$$g(x, \beta) = \frac{\exp\left(x'_{j,t}\beta\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta\right)}$$

and

$$P(x) = \int \frac{\exp\left(x'_{j,t}\beta\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta\right)} f(\beta)\, d\beta.$$

Our goal is to identify the density function $f(\beta)$ in the equation

$$P(x) = \int g(x, \beta)\, f(\beta)\, d\beta. \tag{18}$$

Identification means that a unique $f(\beta)$ solves this equation for all x. We will outline three intermediate, non-primitive conditions that guarantee a model is identified. The appropriate method of proof will depend on the economic model in question. We list them here to show the wide applicability of our mixtures estimator. We then use one of the three methods of proof to show the identification of our main example: the random coefficients logit model.

3.1 Identification with a Finite Number of Types

We begin by considering the identification of a model where the true data generating process is known to have a finite number of types $\beta^{(r)}$, $r = 1, ..., R$. In this case, the true model is

$$P(x) = \sum_{r=1}^{R} g\left(x, \beta^{(r)}\right) f\left(\beta^{(r)}\right). \tag{19}$$

As there are only a finite number of types, the true f (not only the approximation $\bar{f}$ as before) is a probability mass function.

3.1.1 Types are Finite and Known

We first assume that the discrete support points $\beta^{(r)}$ are known to the econometrician and that $P(x)$ is also known to the econometrician. Then, the unknowns in (19) are the R free parameters corresponding to the values of $f\left(\beta^{(r)}\right)$. Suppose that we have R distinct points of support for x, which we label as $x_1, .., x_R$. Simple linear algebra implies that a sufficient condition for identification is that the following R by R matrix is invertible:

$$\begin{bmatrix} g(x_1,\beta_1) & \cdots & g(x_K,\beta_1) \\ \vdots & \ddots & \vdots \\ g(x_1,\beta_K) & \cdots & g(x_K,\beta_K) \end{bmatrix}. \tag{20}$$

If the matrix is nonsingular, then (19) implies that:

$$\begin{bmatrix} f\left(\beta^{(1)}\right) \\ \vdots \\ f\left(\beta^{(R)}\right) \end{bmatrix} = \begin{bmatrix} g\left(x_1,\beta^{(1)}\right) & \cdots & g\left(x_R,\beta^{(1)}\right) \\ \vdots & \ddots & \vdots \\ g\left(x_1,\beta^{(R)}\right) & \cdots & g\left(x_R,\beta^{(R)}\right) \end{bmatrix}^{-1} \begin{bmatrix} P(x_1) \\ \vdots \\ P(x_R) \end{bmatrix}.$$

Now define a class $\mathcal{P} = \left\{ \sum_{r=1}^{R} g\left(x,\beta^{(r)}\right) f\left(\beta^{(r)}\right) : f(\cdot) \geq 0, \sum_{r=1}^{R} f\left(\beta^{(r)}\right) = 1 \right\}$ of all finite mixtures of the family with a finite number R members $\mathcal{G} = \left\{ g\left(x,\beta^{(1)}\right), \ldots, g\left(x,\beta^{(R)}\right) \right\}$, where the points $\beta^{(r)}$ are known. We formalize the invertibility argument.

Theorem 3.1. *Suppose that there exist R real distinct values $x_1, \ldots, x_R$ such that the $R \times R$ matrix (20) is nonsingular. Then, the class $\mathcal{P}$ of all finite mixtures of the finite family $\mathcal{G}$ is identified.*

Proof. See Teicher (1963). Although Teicher (1963)'s work is about a finite mixture of distributions, his proof is general enough to be used for a finite mixture of any suitable functions g. g does not have to be a distribution function. Here we replace Teicher's points in the support of a distribution, x, with our covariates, x. □

3.1.2 Unknown but Finite Types

The above theorem assumes that econometrician knows the points of support and that there are at most a finite number of points of support. A less restrictive approach would be to assume that β lies in an uncountable set, such as some subspace of K dimensional vectors in $\mathbb{R}^K$. We first relax the assumption that econometrician knows the points of support. Therefore, we extend the above result to the case where the family $\mathcal{G}$ is uncountable (e.g., $\mathcal{G} = \left\{ g(x,\beta) : \beta \in \mathcal{B} \subset \mathbb{R}^K \right\}$) but still the class of mixtures $\mathcal{P}$ is generated by a finite subset of the family $\mathcal{G}$.

Theorem 3.2. *For any R distinct values of $\beta \in \mathcal{B}$, suppose that there exist R real distinct values $x_1, \ldots, x_R$ such that the $R \times R$ matrix (20) is nonsingular. Then, the class $\mathcal{P}$ of all finite mixtures of the finite family $\mathcal{G}$ is identified.*

Proof. The proof is a simple modification of Teicher (1963) and is omitted. □

The sufficient condition of Theorem 3.2 means that the family of functions $\mathcal{G}$ constitutes a linearly independent set. When $g(x,\beta)$ is always a distribution function, Teicher (1963) provides a set of condition under which the sufficient condition of Theorem 3.2 holds.

Fox and Gandhi (2007) use similar Teicher linear independence methods. They introduce an intermediate condition for identification that is expressed in terms of the properties of an economic choice model. They show the identification of models where $g(x,\beta)$ is a multinomial choice $j \in J$ or a continuous outcome $y \in \mathbb{R}$. In other words, they do not consider the logit case with parametric distributions for ϵ_{ijt}; all heterogeneity is embedded in the random coefficients. By combing their results on continuous and discrete choices, they prove identification of selection and mixed continuous-discrete models without relying on identification at infinity. Our mixtures estimator could be applied to these selection and mixed continuous-discrete models as well.

3.2 Uncountable Types

Now we relax the second assumption all the number of types to be uncountably infinite. We say the density $f(\beta)$ is identified if there is no other density $\tilde{f}(\beta)$ that satisfies $P(x) = \int g(x,\beta) f(\beta) d\beta$ for all x.

3.2.1 Bounded Completeness

One identification condition is bounded completeness:

Assumption 3.1. *The family $\mathcal{G}$ of measurable function $g(x,\beta)$ in both $x \in \mathcal{X}$ and $\beta \in \mathcal{B}$ is bounded complete in x. Bounded completeness means that if $\int g(x,\beta) h(\beta) d\beta = 0$ for all $x \in \mathcal{X}$ and $h(\cdot)$ is bounded measurable, then $h(\beta) = 0$.*

Theorem 3.3. *Suppose Assumption 3.1 holds. Then, the class of all infinite mixtures $\mathcal{P} = \left\{ \int g(x,\beta) f(\beta) d\beta : M \geq f(\beta) \geq 0, \int f(\beta) = 1 \right\}$ is identified.*

Proof. Identification is equivalent to the nonexistence of any function $\delta(\beta) = \widetilde{f}(\beta) - f(\beta) \neq 0$ such that $\int g(x,\beta) \delta(\beta) d\beta = 0$ for all $x \in \mathcal{X}$, where $\mathcal{X}$ is the support of x. Therefore, identification immediately follows by Assumption 3.1 and the fact that $f(\beta)$ belongs to a class of uniformly bounded density functions. □

That bounded completeness implies identification is nearly tautological, as the two definitions are similar. However, the condition may be useful to the extent bounded completeness has been shown for various classes of functions in the literature. When $g(x,\beta)$ belongs to the exponential family of distributions, completeness is proved Lehmann and Romano (2005, p117). Completeness implies bounded completeness. There are a few other classes of functions that are known to be bounded complete. For example, a location family of absolutely continuous distributions is bounded complete if and only if the characteristic functions of all members of the family are zero-free. Briesch, Chintagunta, and Matzkin (2007) use this condition for identification in nonparametric discrete choice models where the mixture is generated by only the constant term.

We can use a bounded completeness assumption on $g(x,\beta)$ to identify $f(\beta)$ in the logit model. We first identify the distribution of $\eta = x'\beta$ from (19) and then identify the distribution of β from the distribution of η by using the Cramér and Wold (1936) device. The Cramér-Wold device requires that the support of the distribution x be equal to $\mathbb{R}^K$. Discrete covariates are not allowed. Also, applying the Cramér-Wold theorem requires the linear index functional form $x'\beta$; it is not just a convenient simplification.[12]

Any location family of absolutely continuous distributions with compact supports is bounded complete (see Blundell, Chen, and Kristensen (2006) and Lehmann (1986) for related discussions).[13] Other general sufficient conditions for the bounded completeness condition are not available in the literature. Without more results, we need to use alternative identification strategies.

3.2.2 $f(\beta)$ uniquely determined by its moments

Here we propose a third identification condition when $f(\beta)$ satisfies the so-called Carleman condition and g satisfies the single index condition $g(x,\beta) = g(x'\beta)$. A probability measure f satisfying the Carleman condition is uniquely determined by its moments (see Shohat and Tamarkin (1943), p.19). The Carleman condition is weaker than requiring the moment generating function to exist. The main advantage of this alternative identification strategy is that it allows for a general class of $g(x'\beta)$ functions. The component function $g(x'\beta)$ does not have to be a distribution function. We mainly require that $g(x'\beta)$ be continuously differentiable. We do heavily exploit the linear index $x'\beta$.

Assumption 3.2. *(i) The absolute moments of $f(\beta)$, given by $m_l = \int \|\beta\|^l f(\beta)\, d\beta$, are finite for $l \geq 1$ and satisfy the Carleman condition: $\Sigma_{l\geq 1} m_l^{-1/l} = \infty$; (ii) $P(x)$ and $g(x'\beta) \in C^\infty$ (infinitely continuously differentiable) in a neighborhood of $x = 0$; (iii) $g^{(l)}(0)$ is nonzero and finite for all $l \geq 1$ where $g^{(l)}(\cdot)$ denotes the l-th derivative of $g(\cdot)$.*

Assumption 3.2 (iii) restricts the class of $g(x'\beta)$. Some classes of functions satisfies the condition but others do not. For example, if $g(w) = C \cdot \exp(w)$, then Assumption (iii) is trivially satisfied since $g^{(l)}(0) = C$ for all l. If $g(x'\beta)$ is a polynomial function of any finite degree, g does not satisfy the condition since its derivative becomes zero at a certain point. For polynomials, we identify the distribution $f(\beta)$ up to the v-th moment, where v is the order of the polynomial function.

We observe $P(x)$ in the population data and know the function g. We wish to identify the density $f(\beta)$. The general identification argument can be illustrated for the special case where

[12] Using this same, strong support condition and the Cramér-Wold device in a model with only $J = 2$ choices, Ichimura and Thompson (1998) identify the joint distribution of random coefficients and the difference in the product intercepts (they do not rely on the logit model). They require some additional restrictions that likely arise from not imposing the structure of the logit.

[13] A distinct but related definition of bounded completeness is often referred to in the nonparametric IV literature. For an IV setting, d'Haultfoeuille (2007) derives bounded completeness using primitive economic models.

$K = 2$ and so $x'\beta = x_1\beta_1 + x_2\beta_2$. At $x_1 = x_2 = 0$,

$$\left.\frac{\partial P(x)}{\partial x_1}\right|_{x=0} = g^1(0)\int \beta_1 f(\beta)\, d\beta = g^1(0)\, E[\beta_1],$$

where β_1 arises from the chain rule and the expression identifies the mean of β_1 as $P(x)$ is data and $g^1(0)$ is a known constant that does not depend on β. Likewise, $\left.\frac{\partial P(x)}{\partial x_2}\right|_{x=0}/g^1(0)$ equals $E[\beta_2]$, $\left.\frac{\partial^2 P(x)}{\partial x_1 \partial x_2}\right|_{x=0}/g^2(0)$ equals $E[\beta_1\beta_2]$, and $\left.\frac{\partial^2 P(x)}{\partial^2 x_1}\right|_{x=0}/g^2(0)$ equals $E\left[\beta_1^2\right]$. Additional derivatives will identify the other moments of $\beta = (\beta_1, \beta_2)$.

Theorem 3.4. *Let all elements of x be continuous. Suppose Assumption 3.2 holds. Then the class of all infinite mixtures $\mathcal{P} = \left\{\int g(x,\beta) f(\beta)\, d\beta : f(\beta) \geq 0, \int f(\beta) = 1\right\}$ is identified.*

Proof. First we introduce some notation for gradients of arbitrary order, which we need because $f(\beta)$ has a vector of K arguments, β. Let w be a vector of length W. For a function $h(w)$, we denote the $1 \times K^\upsilon$ block vector of υ-th order derivatives as $\nabla^\upsilon h(w)$. $\nabla^\upsilon h(w)$ is defined recursively so that the k-th block of $\nabla^\upsilon h(w)$ is the $1 \times W$ vector $h_k^\upsilon(w) = \partial h_k^{\upsilon-1}(\theta)/\partial w'$, where $h_k^{\upsilon-1}$ is the k-th element of $\nabla^{\upsilon-1} h(w)$. Using a Kronecker product $\otimes$, we can write $\nabla^\upsilon h(w) = \underbrace{\frac{\partial^\upsilon h(w)}{\partial w' \otimes \partial w' \otimes \ldots \otimes \partial w'}}_{\upsilon \text{ Kronecker product of } \partial x'}$.

Take the derivatives with respect to the covariates x of both sides of $P(x) = \int g(x'\beta) f(\beta)\, d\beta$ and evaluate the derivatives at $x = 0$. By Assumption (3.2) (ii), for any $\upsilon = 1, 2, \ldots$ and the chain rule repeatedly applied to the linear index $x'\beta$,

$$\begin{aligned} \nabla^\upsilon P(x)|_{x=0} &= \int \left.g^{(\upsilon)}(x'\beta)\right|_{x=0} \left\{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\right\} f(\beta)\, d\beta \\ &= g^{(\upsilon)}(0) \int \left\{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\right\} f(\beta)\, d\beta. \end{aligned} \tag{21}$$

For each υ there are K^υ equations. Recall g is a known function. Therefore, as long as $g^{(\upsilon)}(0)$ is nonzero and finite for all $\upsilon = 1, 2, \ldots$, we obtain the υ-th moments of $f(\beta)$ for all $\upsilon \geq 1$. Now by Assumption 3.2 (i), $f(\beta)$ satisfies the Carleman condition. Therefore, $f(\beta)$ is identified since a probability measure satisfying the Carleman condition is uniquely determined by its moments. □

3.3 Identification in the Random Coefficients Logit Model

Surprisingly, identification in the random coefficient logit model has never been proved despite its 30 years of use. McFadden and Train (2000) only emphasize the approximating flexibility of the model. Here we use the Theorem 3.4 strategy of determining $f(\beta)$ by its moments. If we make the assumption that $f(\beta)$ is uniquely determined by its moments, than the only further condition we need to verify is that the g function for the logit model has nonzero derivatives at 0. We also require that all elements of x be continuous, as (21) requires derivatives with respect to x.

First consider the case of $J \geq 3$ choices. The multinomial logit with random coefficients model with $J \geq 3$ is identified if $f(\beta)$ satisfies Assumption 3.2 (i). Let $x_j = 0$ for all $j \neq 1$. Then the choice probability of alternative 1 given β is

$$g_1\left(x_1'\beta, \ldots, x_J'\beta\right) = g_1\left(x_1'\beta, 0, \ldots, 0\right) = \frac{\exp\left(x_1'\beta\right)}{J - 1 + \exp\left(x_1'\beta\right)}.$$

Algebra can verify that $g_1^{(l)}(0) \neq 0$ for all l. Thus $f(\beta)$ is identified by Theorem 3.4. This result is new despite the random coefficient logits' wide use in the empirical literature. Note also its simplicity: from Theorem 3.4, we need only to check for non-zero derivatives of $g(a)$ at $a = 0$. This technique can be applied to show identification of many other differentiable economic models.

Note the identification of the logit mixtures model occurs by varying the linear index $x_1'\beta$ around a neighborhood of 0. This is a very local form of identification, and is much weaker on the data than identification arguments that relay on identification at infinity, such as Lewbel (2000). On the other hand, Lewbel uses a large-support "special regressor" to avoid our assumption of the independence of x and ϵ, does not require that all elements of x_1 be continuous, and does not use the logit, so the two sets of assumptions are non-nested.

Now consider the somewhat special case of $J = 2$ choices, where x represents the difference between the two products' characteristics: $x = x_1 - x_2$. The probability of purchasing product 1 is

$$g\left(x'\beta\right) = \frac{\exp\left(x'\beta\right)}{1 + \exp\left(x'\beta\right)}.$$

Unfortunately, algebra shows that $g^{(l)}(0) \neq 0$ when l is odd and $g^{(l)}(0) = 0$ when l is even. The zero derivatives mean the $J = 2$ case is degenerate so we will need to impose that the true density of β generates statistically independent random variables. In other words, we show the logit mixtures model with $f(\beta) = \prod_{k=1}^{K} f_k(\beta_k)$ (an independent multivariate distribution) that also satisfies Assumption 3.2 (i) is identified. We identify the odd moments of β from (21) and the nonzero derivatives of g, but since $f(\beta) = \prod_{k=1}^{K} f_k(\beta_k)$, we also identify any even moments. To see this for an example, from (21) we can obtain two odd moments such as $E[\beta_1\beta_2^2]$ and $E[\beta_1]$. As β_1 and β_2 are independent, we also obtain the even moment $E[\beta_2^2] = E[\beta_1\beta_2^2]/E[\beta_1]$.[14] Similar arguments show that we can identify other even moments.

4 Large Sample Theory with Finite Types

Although the logit model is our primary example in this paper, the following asymptotic results are derived under general mixtures models. Our main contribution is to take the boundary value problem seriously: some of the weights can be zero in the true data generating process. This

[14] When $E[\beta_1] = 0$, one can use $E[\beta_1^3\beta_2^2]$ and $E[\beta_1^3]$ to obtain $E[\beta_2^2] = E[\beta_1^3\beta_2^2]/E[\beta_1^3]$.

complicates computing confidence regions. However, the more conservative confidence regions from OLS will still have correct coverage, even if the parameter is on the boundary. Keep in mind that one can ignore the boundary problem and use the OLS confidence regions, if only to save on programming time.

4.1 Asymptotic Distribution of Inequality Constrained LS with Finite Types

Here we derive the asymptotic distribution of the estimator for $f(\beta)$ when the data are observed choices across different markets. We let $f = \left(f\left(\beta^{(1)}\right), \ldots, f\left(\beta^{(R)}\right)\right)'$ and f_0 denote the true f. In other words, the true data generating process has a finite number of types and those types are known. $f(\beta)$ is a probability mass function.

The asymptotic distribution is nonstandard because the estimable weights, $f(\beta)$, satisfy the equality and inequality constraints $f(\beta) \geq 0$ and $\sum_{r=1}^{R} f(\beta^{(r)}) = 1$. The asymptotic distribution is not pivotal in the sense that the asymptotic distribution depends on whether the true $f_0(\beta)$'s are on the boundary or not. For example, if $f_0(\beta^{(1)}) = 0$, then $\widehat{f}(\beta^{(1)}) - f_0(\beta^{(1)}) = \widehat{f}(\beta^{(1)}) \geq 0$ and so the distribution of $\widehat{f}(\beta^{(1)}) - f_0(\beta^{(1)})$ can only take non-negative values while the distribution can take both positive and negative values when $f_0(\beta^{(1)}) > 0$. In many applications, this boundary problem has been ignored because a researcher believes that the true parameters are not on the boundary. In our Monte Carlo experiments, we typically find that only small portions of the generated bases have positive mass. To some degree these Monte Carlo findings are comforting as our estimator does not invent mass points where they are not needed. Therefore, we derive the asymptotic distribution of our estimator while explicitly considering this boundary problem.

Due to Andrews (1999, 2002), and as shown below, the asymptotic distribution is (if we ignore the constraint $\sum_{r=1}^{R} f(\beta^{(r)}) = 1$ for now)

$$\sqrt{T}\left(\widehat{f} - f_0\right) \underset{d}{\longrightarrow} \begin{bmatrix} 1(f_0(1) = 0)\mathcal{Z}_1^+ + 1(f_0(1) > 0)\mathcal{Z}_1 \\ 1(f_0(2) = 0)\mathcal{Z}_2^+ + 1(f_0(2) > 0)\mathcal{Z}_2 \\ \vdots \\ 1(f_0(R) = 0)\mathcal{Z}_R^+ + 1(f_0(R) > 0)\mathcal{Z}_R \end{bmatrix}, \tag{22}$$

where the random vector $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_R)'$ follows the asymptotic distribution of OLS without constraints and $\mathcal{Z}_r^+$ denotes a half normal distribution that takes only the non-negative values of $\mathcal{Z}_r$ for $r = 1, \ldots, R$. This result is intuitive. When $f_0(\cdot)$ is not equal to zero (and so is not on the boundary), then $\widehat{f}(\cdot) - f_0(\cdot)$ can be both positive or negative and the usual normal distribution given by the central limit theorem applies. Now when $f_0(\cdot) = 0$ (on the boundary), then $\widehat{f}(\cdot) - f_0(\cdot)$ can be only positive or zero since $\widehat{f}(\cdot) - f_0(\cdot) \geq 0$. Therefore, $\widehat{f}(\cdot) - f_0(\cdot)$ will follow a half normal distribution and take non-negative values only.

We now formalize this argument to obtain the asymptotic distribution of the ICLS estimator.

Denote

$$d_{r,j,t} = g\left(x_t, \beta^{(r)}\right) = g\left(x'_{j,t}\beta^{(r)}, \ldots, x'_{J,t}\beta^{(r)}\right)$$

where the notation j for the J alternatives is for the multinomial choice example, not a general mixtures model. The market or observation index is t. For the logit,

$$d_{r,j,t} = \left(\exp\left(x'_{j,t}\beta^{(r)}\right) \Big/ \sum_{j=1}^{J} \exp\left(x'_{j',t}\beta^{(r)}\right)\right).$$

Now let $D(j)$ be an $T \times R$ matrix that stacks $d_{r,j,t}$ systematically such that r-th column of $D(j)$ stacks $d_{r,j,t}$ given r. Also let $d_{j,t} = (d_{1,j,t}, \ldots, d_{R,j,t})'$ and denote $\Xi = \text{plim}_{T\to\infty} \frac{\sum_{j=1}^{J-1}\sum_t d_{j,t}d'_{j,t}}{T}$ and [15]

$$\Omega = \text{plim}_{T\to\infty} \frac{\sum_{1\le j\neq j'\le J-1}\sum_t d_{j,t}d'_{j',t}e_{jt}e_{j't}}{T} \quad \text{where } e_{jt} = \widehat{P}(j,t) - P\left(j|x_t, f\right).$$

By construction, Ω is robust to the possible correlation and heteroskedasticity between e_{jt}'s across different alternatives and is also robust to possible heteroskedasticity across different markets. Under zero correlation and homoskedasticity across different alternatives and homoskedasticity across different markets, Ω simplifies to $(J-1)E[e_{jt}^2]\Xi^{-1}$. Also, under zero correlation but heteroskedasticity across different alternatives and heteroskedasticity across different markets, Ω simplifies to $\text{plim}_{T\to\infty} \frac{\sum_{j=1}^{J-1}\sum_t d_{j,t}d'_{j,t}e_{jt}^2}{T}$. Finally, let $\widehat{P}(j)$ denote a $T \times 1$ vector that stacks $\widehat{P}(j,t)$ in the same manner with $D(j)$.

Then, under some regularity conditions, following Andrews (1999, 2002), we obtain the asymptotic distribution of $\widehat{f}$ as:

$$\sqrt{T}\left(\widehat{f} - f_0\right) \underset{d}{\longrightarrow} \lambda^* \tag{23}$$

where

$$\begin{aligned} \lambda^* &= \arg\inf_{\lambda\in\Lambda} q(\lambda),\ q(\lambda) = (\lambda - \mathcal{Z})'\Xi(\lambda - \mathcal{Z}) \\ \mathcal{Z} &\sim \mathcal{N}\left(0, \Xi^{-1}\Omega\Xi^{-1}\right) \\ \Lambda &= \left\{\lambda \in \mathbb{R}^R : I_b\lambda \geq \mathbf{0}\right\} \end{aligned} \tag{24}$$

and I_b denotes the submatrix of an identity matrix I that consists of the rows of I such that $I_b f_0 = \mathbf{0}$. The optimization problem in (24) is solved once for each realization z of the normal random variable $\mathcal{Z}$. $\mathcal{Z}$ is a distribution of a random variable, not a PDF or CDF. By construction of the set Λ, when a constraint is binding for $f_0(\beta^{(r)})$, it restricts the values that the corresponding λ_r takes, such as $\lambda_r \geq 0$. λ_r is unrestricted if no constraint is binding for $f_0(\beta^{(r)})$. Therefore, the solution of the minimization problem $\lambda^* = \arg\inf_{\lambda\in\Lambda}(\lambda - \mathcal{Z})'\Xi(\lambda - \mathcal{Z})$ will pick the whole

[15] We have at most $J-1$ independent market shares or choice probabilities because $\sum_{j=1}^{J} e_{jt} = \sum_{j=1}^{J}(\widehat{P}(j,t) - P\left(j|x_t, f\right)) = 1 - 1 = 0$.

distribution $\mathcal{Z}$ when it is unrestricted and half of the distribution $\mathcal{Z}$ when it is restricted.

Even though the distribution of $\sqrt{T}\left(\widehat{f}-f_0\right)$ is non-standard and not pivotal (it depends on the true value f_0 through I_b and so Λ), it can be easily simulated since λ^* is the solution to a quadratic programming problem for each realization of $\mathcal{Z}$. If the true model has finite types, there is no need for simulation as the closed form solution of (24) is given by (22).

A test statistic can be easily constructed[16] from such a simulated distribution because consistent estimators of Ξ and Ω can be obtained from their sample analogues. If one wants to construct a confidence interval or confidence set, one should use a resampling method such as subsampling because the asymptotic distribution depends on the population parameter values, which are not known. Interestingly, Andrews and Guggenberger (2005) show that subsampling has poor coverage in this case and the usual confidence interval ignoring the boundary problem (use the critical value from the OLS distribution) has a negligible size distortion. The reason that subsampling has poor coverage is the greater variability of the statistic in one of the chosen subsamples than in the full sample. In other words, a subset of the data is affected by the boundary problem more than the full sample and therefore subsampling does not approximate the true distribution well. Following Andrews and Guggenberger, one suggestion that also eases programming is that users employ the standard OLS confidence regions. These may be too conservative, but will often have at least 95% coverage.

$\mathcal{Z}$ is the asymptotic distribution of $\sqrt{T}\left(\widetilde{f}-f_0\right)$ where $\widetilde{f}$ denotes the unconstrained OLS estimator such that $\widetilde{f}=\left(\sum_{j=1}^{J}D(j)'D(j)\right)^{-1}\left(\sum_{j=1}^{J}D(j)'\widehat{P}(j)\right)$ and thus $\Xi^{-1}\Omega\Xi^{-1}$ is the heteroskedastic robust asymptotic variance-covariance matrix of $\widetilde{f}$, which reduces to $\sigma_e^2\Xi^{-1}$ with $\sigma_e^2=E[e_{jt}^2]$ under no correlation across different alternatives and under homoskedasticity across both different alternatives and markets.

The asymptotic distribution in (23) can be also used when we add the equality constraint $\sum_{r=1}^{R}f(\beta^{(r)})=1$. Simply redefine $\widetilde{P}(j,t)\equiv\widehat{P}(j,t)-d_{1,j,t}$ and $\widetilde{d}_{k,j,t}\equiv d_{k,j,t}-d_{1,j,t}$ for $k=2,\ldots,R$ and construct $D(j)$ and $\widehat{P}(j)$ based on $\widetilde{P}(j,t)$ and $\widetilde{d}_{k,j,t}$. Then, let $f=(f(\beta^{(2)}),\ldots,f(\beta^{(R)}))'$ and $\Lambda=\{\lambda\in\mathbb{R}^{R-1}:I_b\lambda\geq\mathbf{0}\}$ without loss of generality. Alternatively we can replace Λ in (23) with $\Lambda=\{\lambda\in\mathbb{R}^{R}:\lambda_1+\lambda_2+\ldots+\lambda_R=0,\,I_b\lambda\geq\mathbf{0}\}$ if we keep the original specification of the model and add the $\sum_{r=1}^{R}f(\beta^{(r)})=1$ equality constraint. In the above construction of Λ, we have $\sum_{r=1}^{R}\lambda_R=0$, not $\sum_{r=1}^{R}\lambda_R=1$. It is not a mistake. This is because λ represents the difference, $\widehat{f}-f_0$, not $\widehat{f}$ itself. We, therefore, have $\sum_{r=1}^{R}\lambda_R=\sum_{r=1}^{R}\left(\widehat{f}(\beta^{(r)})-f_0(\beta^{(r)})\right)=\sum_{r=1}^{R}\widehat{f}(\beta^{(r)})-\sum_{r=1}^{R}f_0(\beta^{(r)}=1-1=0$.

[16] To test a hypothesized value of f_0, the relevant standard error can be obtained by plugging in the hypothesized value of f_0 in the asymptotic distribution.

4.2 Nonlinear Extension: Inequality Constrained NLS

Now we consider the location and scale model of Section 2.4 where the inequality constrained, nonlinear least squares (ICNLS) estimator is given by (16). The asymptotics developed here can be used in other cases where the model contain a subset of parameters without mixtures. Characterizing the asymptotic distribution of these parameters is of separate interest. For example, one can estimate price elasticities from the random utility model and use the following asymptotic result to obtain the standard errors of such estimators. Here we will see that the asymptotic distributions of the location and the scale parameters (and other parameters without mixtures) depend on the asymptotic distribution of the frequency parameter estimators.

Again using general notation for a discrete choice model with J alternatives, we denote

$$d_{r,j,t}(\mu,\sigma) = g\left(\sum_{k=1}^{K} x_{k,1,t}(\mu_k + \sigma_k\beta_k^{(r)}), \ldots, \sum_{k=1}^{K} x_{k,J,t}(\mu_k + \sigma_k\beta_k^{(r)})\right)$$

for alternative j , where j's attributes are $x_{j,t} = (x_{1,j,t}, \ldots, x_{K,j,t})'$ in market t. Also, μ and σ are the collection of means and standard errors: $\mu = (\mu_1, \ldots, \mu_K)'$ and $\sigma = (\sigma_1, \ldots, \sigma_K)'$. In the example of the logit model, we have

$$d_{r,j,t}(\mu,\sigma) = \exp\left(\sum_{k=1}^{K} x_{k,j,t}\left(\mu_k + \sigma_k\right)\beta_k^{(r)}\right) \Bigg/ \sum_{j'=1}^{J} \exp\left(\sum_{k=1}^{K} x_{k,j',t}\left(\mu_k + \sigma_k\beta_k^{(r)}\right)\right).$$

We derive the asymptotic distributions of the estimator $(\widehat{\mu}', \widehat{\sigma}', \widehat{f}')'$ obtained from (16). We also let $f(r) = f(\beta^{(r)})$ and $h(j,t;\mu,\sigma,f) = \sum_{r=1}^{R} f(r) d_{r,j,t}(\mu,\sigma)$. Then (16) can be rewritten as an M-estimator based on the following moments for $j = 1, \ldots, J$:

$$E\left[g_1(j,\cdot;\mu,\sigma,f)\right] \equiv E\left[h_1(j,\cdot;\mu,\sigma,f)(h(j,\cdot;\mu,\sigma,f) - \widehat{P}(j,t))\right] = 0 \tag{25}$$

$$E\left[g_2(j,\cdot;\mu,\sigma,f)\right] \equiv E\left[h_2(j,\cdot;\mu,\sigma,f)(h(j,\cdot;\mu,\sigma,f) - \widehat{P}(j,t))\right] = 0 \tag{26}$$

where $h_1(j,\cdot;\mu,\sigma,f) = \frac{\partial h(j,\cdot;\mu,\sigma,f)}{\partial(\mu',\sigma')'}$ and $h_2(j,\cdot;\mu,\sigma,f) = \frac{\partial h(j,\cdot;\mu,\sigma,f)}{\partial f}$. If the model contains other parameters without mixtures, say ϑ, to be estimated, simply add an additional set of moment conditions, $E\left[g_3(j,\cdot;\mu,\sigma,\vartheta,f)\right] \equiv E\left[h_3(j,\cdot;\mu,\sigma,\vartheta,f)(h(j,\cdot;\mu,\sigma,\vartheta,f) - \widehat{P}(j,t))\right] = 0$ where $h_3(j,\cdot;\cdot) = \frac{\partial h(j,\cdot;\mu,\sigma,\vartheta,,f)}{\partial\vartheta}$ and consider the M-estimation problem of the following moments conditions for $j = 1, \ldots, 1$:

$$\begin{aligned}
E\left[g_1(j,\cdot;\mu,\sigma,\vartheta,f)\right] &\equiv E\left[h_1(j,\cdot;\mu,\sigma,\vartheta,f)(h(j,\cdot;\mu,\sigma,\vartheta,f) - \widehat{P}(j,t))\right] = 0 \\
E\left[g_2(j,\cdot;\mu,\sigma,\vartheta,f)\right] &\equiv E\left[h_2(j,\cdot;\mu,\sigma,\vartheta,f)(h(j,\cdot;\mu,\sigma,\vartheta,f) - \widehat{P}(j,t))\right] = 0 \\
E\left[g_3(j,\cdot;\mu,\sigma,\vartheta,f)\right] &\equiv E\left[h_3(j,\cdot;\mu,\sigma,\vartheta,f)(h(j,\cdot;\mu,\sigma,\vartheta,f) - \widehat{P}(j,t))\right] = 0
\end{aligned}$$

where $h_1(j,\cdot;\mu,\sigma,\vartheta,f) = \frac{\partial h(j,\cdot;\mu,\sigma,\vartheta,f)}{\partial(\mu',\sigma')'}$ and $h_2(j,\cdot;\mu,\sigma,\vartheta,f) = \frac{\partial h(j,\cdot;\mu,\sigma,\vartheta,f)}{\partial f}$.

Let $\widehat{\theta} = (\widehat{\mu}', \widehat{\sigma}', \widehat{f}')'$ denote the ICNLS estimator in (16) and $\theta_0 = (\mu_0', \sigma_0', f_0')'$ denote the true parameter. Then, under a set of regularity conditions (Appendix A), we can show that $\widehat{\theta}$ is a consistent estimator of θ_0. Now we turn to the asymptotic distribution of $\widehat{\theta}$. Here we restrict our discussion to the asymptotic distribution of $\widehat{\theta}$ because it will have a nonstandard distribution due to the boundary problem of the frequency parameters.

We first derive the asymptotic distribution of $\widehat{\theta}$ assuming θ_0 is an interior point of the parameter space (no boundary problem). The asymptotic distribution of a nonlinear least squares estimator is well known in the literature (Amemiya (1983)). Under some regularity conditions, we have

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \underset{d}{\longrightarrow} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$$

where

$$\begin{aligned} H &= \sum\nolimits_{j=1}^{J-1} E\begin{bmatrix} h_1(j,\cdot;\theta_0)h_1(j,\cdot;\theta_0)' & h_1(j,\cdot;\theta_0)h_2(j,\cdot;\theta_0)' \\ h_2(j,\cdot;\theta_0)h_1(j,\cdot;\theta_0)' & h_2(j,\cdot;\theta_0)h_2(j,\cdot;\theta_0)' \end{bmatrix} \text{ and} \\ \Sigma &= \sum\nolimits_{1\le j\neq j'\le J-1} E\left[\begin{pmatrix} h_1(j,\cdot;\theta_0)h_1(j',\cdot;\theta_0)' & h_1(j,\cdot;\theta_0)h_2(j',\cdot;\theta_0)' \\ h_2(j,\cdot;\theta_0)h_1(j',\cdot;\theta_0)' & h_2(j,\cdot;\theta_0)h_2(j',\cdot;\theta_0)' \end{pmatrix} e_{jt}e_{j't}\right]. \end{aligned}$$

Σ is robust to the possible correlation and heteroskedasticity between e_{jt}'s across different alternatives and is also robust to possible heteroskedasticity across different markets.

We partition $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}' & H_{22} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix}$ such that $H_{11}, \Sigma_{11} \in \mathbb{R}^{2K\times 2K}$, $H_{12}, \Sigma_{12} \in \mathbb{R}^{2K\times R}$, and $H_{22}, \Sigma_{22} \in \mathbb{R}^{R\times R}$. We also let $\mathcal{V} = \begin{bmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \end{bmatrix} \sim \mathcal{N}(0,\Sigma)$ with $\mathcal{V}_1 \in \mathbb{R}^{2K}$ and $\mathcal{V}_2 \in \mathbb{R}^R$. Here the subscript "1"corresponds to the location and scale parameters and "2"denotes the type frequency parameters.

Following Andrews (1999, 2002), we derive the asymptotic distribution when f_0 is possibly on the boundary:[17]

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \underset{d}{\longrightarrow} \lambda^{**} \tag{27}$$

and $\lambda^{**} = \arg\inf_{\lambda\in\Lambda\equiv\Lambda_1\times\Lambda_2} q(\lambda)$ where

$$\begin{aligned} q(\lambda) &= (\lambda - \mathcal{Z})'H(\lambda - \mathcal{Z}) \\ \mathcal{Z} &\sim \mathcal{N}(0, H^{-1}\Sigma H^{-1}) \\ \Lambda_1 &= \left\{\left(\lambda_1', \lambda_2'\right)' \in \mathbb{R}^{2K}\right\}, \Lambda_f = \left\{\lambda_f \in \mathbb{R}^R : \sum\nolimits_{r=1}^{R} \lambda_{f,r} = 0, I_b\lambda_f' \ge 0\right\}, \end{aligned}$$

[17]The true σ_0 is assumed to be positive and so it does not create an additional boundary problem.

and I_b denotes the submatrix of an identity matrix I that consists of the rows of I such that $I_b f_0 = \mathbf{0}$. Noting that the location and the scale parameter (μ, σ) do not create a boundary problem, we can decompose the asymptotic distribution $(\widehat{\mu}', \widehat{\sigma}')'$ and $\widehat{f}$ using partitioned matrix inversion:

$$\sqrt{T}\left(\widehat{f} - f_0\right) \underset{d}{\longrightarrow} \lambda_f^{**}$$

and $\lambda_f^{**} = \arg\inf_{\lambda_f \in \Lambda_f} q_f(\lambda_f)$ where

$$\begin{aligned} q_f(\lambda_f) &= (\lambda_f - \mathcal{Z}_f)' \left(CH^{-1}C'\right)^{-1} (\lambda_f - \mathcal{Z}_f) \\ \mathcal{Z}_f &= C\mathcal{Z},\ C = \left[\begin{array}{cc} \mathbf{0} & I \end{array}\right] \in \mathbb{R}^{R \times (2K+R)} \end{aligned}$$

and

$$\sqrt{T}\left(\left(\widehat{\mu}', \widehat{\sigma}'\right)' - \left(\mu_0', \sigma_0'\right)'\right) \underset{d}{\longrightarrow} H_{11}^{-1}\mathcal{V}_1 - H_{11}^{-1}H_{12}\lambda_f^{**} \tag{28}$$

and the convergence of $\sqrt{T}\left(\widehat{f} - f_0\right)$ and $\sqrt{T}\left((\widehat{\mu}', \widehat{\sigma}')' - (\mu_0', \sigma_0')'\right)$ holds jointly. We conclude that, as before, $\sqrt{T}\left(\widehat{f}(\cdot) - f_0(\cdot)\right)$ follows a half-normal or normal distribution depending on whether the true parameter is on the boundary or not. $\sqrt{T}\left((\widehat{\mu}', \widehat{\sigma}')' - (\mu_0', \sigma_0')'\right)$ also follows a nonstandard distribution since it depends on λ_f^{**}, the distribution of the frequency parameter estimators.

We further note that

$$\mathcal{Z}_f = C\mathcal{Z} = H_{22}^{-1}\mathcal{V}_2 + H_{22}^{-1}H_{21}\left(H_{11} - H_{12}H_{22}^{-1}H_{21}\right)^{-1}\left(H_{12}H_{22}^{-1}\mathcal{V}_2 - \mathcal{V}_1\right).$$

A test statistic can be constructed from these asymptotic results since consistent estimators of H and Σ are easily obtained from their sample analogues.

4.3 Optimal GMM Estimator Combining J Market Shares

Until now we have disregarded the fact that we can estimate the full set of parameters $\widehat{\theta}$ for each $j = 1, \ldots, J$. In other words, we have J sets of point estimates $\widehat{\theta}$. A natural question is how to combine these $\widehat{\theta}$'s optimally. The optimally weighted estimator uses two-step GMM. We have moment conditions (25) and (26) for $j = 1, \ldots, J-1$.[18] Now let $g(t; \mu, \sigma, f)$ collect $g(j, t; \mu, \sigma, f)$ for $J-1$ market shares, meaning $g(t; \mu, \sigma, f) = \left\{g(1, t; \cdot)', \ldots, g(J-1, t; \cdot)'\right\}'$. Also let $g(j, t; \mu, \sigma, f) = (g_1(j, t; \cdot)', g_2(j, t; \cdot)')'$. Then, the GMM estimator is given by

$$\widehat{\theta}(A) = \arg\min \left(\frac{1}{T}\sum_t g(t; \mu, \sigma, f)\right)' A \left(\frac{1}{T}\sum_t g(t; \mu, \sigma, f)\right).$$

[18] Since we have only $J-1$ independent alternatives out of J.

The optimal two-step GMM estimator is $\widehat{\theta}(\widehat{A})$ with the weighting matrix $\widehat{A} = \left(\frac{1}{T} \sum_t g(t; \widehat{\theta}(I)) g(t; \widehat{\theta}(I))' \right)^{-1}$ where $\widehat{\theta}(I)$ denotes the first-step GMM estimator with $A = I \in \mathbb{R}^{(J-1)(R+2K) \times (J-1)(R+2K)}$. The original ICNLS estimator in (16) is equal to the first-step GMM estimator with the identify weighting matrix. The asymptotic distribution of the optimal GMM estimator is provided in Appendix B.

5 Large Sample Theory with Uncountable Types

In this section, we present the large sample theory of our estimator when the number of types is possibly infinite. We derive only rates of convergence, both for the approximation of market shares and for the approximation of the underlying distribution of types $F(\beta)$. To simplify our notation, let $X_t = \left(X'_{1,t}, \ldots, X'_{J,t} \right)'$ and denote by $\mathcal{X} \equiv \mathcal{X}_1 \times \ldots \times \mathcal{X}_J$ the support of the distribution of X_t.

Our starting point is McFadden and Train (2000). They prove (Theorem 1) that a random coefficients logit model can approximate any random utility model with linear indices $x'\beta$ in utilities. Therefore, due to McFadden and Train, for arbitrary small η, we can let

$$P(j, x_t, f_0) \equiv P(j,t) = \sum\nolimits_{r=1}^{\mathbf{R}} f(r) \left(\frac{\exp\left(x'_{j,t} \beta^{(r)} \right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t} \beta^{(r)} \right)} \right) + \eta \text{ for any } x_t = \left(x'_{1,t}, \ldots, x'_{J,t} \right)' \in \mathcal{X}.$$

where $\mathbf{R}$ could be very large or infinite.

We want to approximate the true market share $P(j,t)$ using the series approximation of the observed market shares $\widehat{P}(j, x_t) \equiv \widehat{P}(j,t)$ by our basis functions:

$$\widehat{P}(j, x_t) = \sum_{r=1}^{R} \bar{f}(r) \, d(r, j, x_t) + e_{jt} \tag{29}$$

where $e_{jt} = \widehat{P}(j, x_t) - P(j, x_t)$ and $d(r, j, x_t) = \frac{\exp\left(x'_{j,t} \beta^{(r)} \right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t} \beta^{(r)} \right)}$. We introduce additional simplifying notation. With some duplication of notation, let

$$d_f(j, x_t) = \sum_{r=1}^{R} \bar{f}(r) \, d(r, j, x_t).$$

Recall that we require $\bar{f}(r) \geq 0$ and $\sum_{r=1}^{R} \bar{f}(r) = 1$ as the $\bar{f}(r)$'s are type frequencies in our approximation. We can approximate $P(j,t)$ arbitrarily well using our basis functions $d(r, j, x_t)$ noting $d_f(j, x_t) - P(j, x_t, f_0) \to 0$ as $R \to \mathbf{R}$ by construction of $P(j, x_t, f_0)$ and $d_f(j, x_t)$. Therefore,

the least squares estimator is

$$\widehat{f} = \text{argmin}_{f \in \Delta^R} \sum_{j=1}^{J-1} \frac{1}{T} \sum_{t=1}^{T} \left\{ \widehat{P}(j, x_t) - d_f(j, x_t) \right\}^2 \tag{30}$$

where $\Delta^R = \left\{ f = \left(\bar{f}(1), \ldots, \bar{f}(R)\right)' \in \mathbb{R}^R : \bar{f}(r) \geq 0 \text{ and } \sum_{r=1}^{R} \bar{f}(r) = 1 \right\}$.

In what follows, even though our notation is specific to the logit example, we derive the asymptotic result for any basis functions that satisfy a set of conditions. We let $\|g\|_T = \sqrt{\frac{1}{T}\sum_{t=1}^{T} g^2(x_t)}$, $\|g\|_0^2 = \int_{\mathcal{X}} g^2(x) d\mu(x)$ (the norm in L_2), $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)|$ for any function $g : \mathcal{X} \to \mathbb{R}$ where μ denotes the distribution of X.

Assumption 5.1. *(i) $\{e_{jt}, \ldots, e_{Jt}\}$ are independent across $t = 1, \ldots, T$; (ii) $E[\varepsilon_{jt}|X_1, \ldots, X_T] = 0$ and $E[\varepsilon_{jt}^2|X_1, \ldots, X_T]$ is bounded for all $j = 1, \ldots, J$; (iii) $\{X_{j,t}, \ldots, X_{J,t}\}$ are iid across $t = 1, \ldots, T$.*

Assumption 5.2. *(i) $\|P(j, x, f_0)\|_\infty \leq \zeta_0$ and $\|d(r, j, x)\|_\infty \leq \zeta_0$ for all $r \leq R$ uniformly and for all j; (ii) the $R \times R$ matrix $\left(E\left[d(r, j, \cdot)d(r', j, \cdot)\right]\right)_{1 \leq r, r' \leq R}$ is positive definite and the smallest eigenvalue of $\left(E\left[d(r, j, \cdot)d(r', j, \cdot)\right]\right)_{1 \leq r, r' \leq R}$ $(\equiv \xi_{\min}(R))$ is bounded away from zero uniformly in R and for all j.*

Note that Assumption 5.1 is a standard assumption. In Assumption 5.1 (i), we do not impose the iid condition. $\{e_{jt}, \ldots, e_{Jt}\}$ are not identically distributed to allow for heteroskedasticity. For the logit case, we satisfy Assumption 5.2 (i) trivially since

$$d(r, j, x) = \frac{\exp\left(x_j' \beta^{(r)}\right)}{\sum_{j'=1}^{J} \exp\left(x_{j'}' \beta^{(r)}\right)} \leq 1 \text{ uniformly over } \beta \text{ and } x$$

and $P(j, x, f_0) \leq 1$ uniformly by construction. We can also satisfy Assumption 5.2 (ii) by constructing $d(r, \cdot)$ $(1 \leq r \leq R)$ sequentially with careful choices of β's. In practice, we have only to make the $(R \times R)$ matrix $\left(\frac{1}{T}\sum_{t=1}^{T} d(r, j, x_t) d(r', j, x_t)\right)_{1 \leq r, r' \leq R}$ nonsingular.

Under these regularity conditions, the rate of convergence of the market share approximation is

Theorem 5.1. *Suppose Assumptions 5.1 and 5.2 hold. Further suppose $\frac{R^2 \log R}{T} \to 0$. Then, we have*

$$\sum_{j=1}^{J-1} \|d_f(j, x_t) - P(j, x_t, f_0)\|_T^2 = O_{\mathbb{P}}\left(\frac{J \cdot R \cdot \log(RT)}{T} + \inf_{f \in \Delta^R} \sum_{j=1}^{J-1} \|d_f(j, x_t) - P(j, x_t, f_0)\|_T^2 \right).$$

The first term is reasonably fast. We can also obtain the convergence rate results as Theorem

5.2 if we are willing to assume that the number of types does not grow too fast compared to the number of markets (sample size) in the following sense:

Assumption 5.3. *There exist* $f^* \in \Delta^R$ *and a positive constant* C_* *such that* $\frac{1}{J-1}\sum_{j=1}^{J-1} ||d_{f^*}(j,x) - P(j,x,f_0)||_\infty^2 \leq C_* \eta_T^2(R) R$ *for a sequence* $\eta_T(R) \underset{T\to\infty}{\longrightarrow} 0$.

In other words, the number of types required to explain heterogeneous consumer behaviors in markets are not that large or finite. This assumption is generally acceptable. For example, we can identify only up to $J!$ types with J different alternative choices with finite alternatives.

Theorem 5.2. *Suppose Assumptions 5.1, 5.2, and 5.3 hold for the RUM model. Further suppose* $\eta_T = \sqrt{A\frac{\log(RT)}{T}}$ *with a constant* $A > 8/3$ *and* $\frac{R^2 \log R}{T} \to 0$. *Then, we have*

$$\begin{aligned} \sum_{j=1}^{J-1} \left\| d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0) \right\|_T^2 &= O_{\mathbb{P}}\left(\frac{J \cdot R \log(RT)}{T} \right) \text{ and} \\ \sum_{r=1}^{R} \left| \widehat{f}(r) - f^*(r) \right| &= O_{\mathbb{P}}\left(R\sqrt{\frac{\log(RT)}{T}} \right). \end{aligned}$$

Theorem 5.2 establishes the convergence rates for both the distance between the true market share and the approximated market share as well as the L_1 distance between $\widehat{f}$ and f^*. The L_1 metric for the discrepancy between $\widehat{f}$ and f^* is natural because $\widehat{f}$ and f^* are frequency parameters.

Denote by $F_0(\beta)$ the distribution of β and let $f_0(\beta)$ be its density function. Also let $\mathcal{B}$ be the support (or compact subset of the support) of $F_0(\beta)$. One would estimate the distribution of the random coefficients using the following empirical distribution

$$\widehat{F}_T(\beta) = \sum_{r=1}^{R} \widehat{f}(r) 1\left\{\beta^{(r)} \leq \beta\right\}.$$

Now define a pseudo true distribution such that $F_R(\beta) = \sum_{r=1}^{R} f^*(\beta) 1\left\{\beta^{(r)} \leq \beta\right\}$.

Then,

$$\begin{aligned} \sup_{\beta\in\mathcal{B}} \left| \widehat{F}_T(\beta) - F_R(\beta) \right| &= \sup_{\beta\in\mathcal{B}} \left| \sum_{r=1}^{R} \widehat{f}(r) 1\left\{\beta^{(r)} \leq \beta\right\} - \sum_{r=1}^{R} f^*(r) 1\left\{\beta^{(r)} \leq \beta\right\} \right| \\ &= \sup_{\beta\in\mathcal{B}} \left| \sum_{r=1}^{R} \left(\widehat{f}(r) - f^*(r) \right) 1\left\{\beta^{(r)} \leq \beta\right\} \right| \\ &\leq \sum_{r=1}^{R} \left| \widehat{f}(r) - f^*(r) \right| = O_{\mathbb{P}}\left(R\sqrt{\frac{\log(RT)}{T}} \right) \end{aligned}$$

where the last inequality holds by the triangle inequality and the last equality holds by Theorem 5.2. Noting that $F_R(\beta) \underset{R\to\infty}{\longrightarrow} F_0(\beta)$ a.e. in $\beta \in \mathcal{B}$ as long as $F_0(\beta)$ is Lebesgue integrable and $f_0(\beta)$

is uniformly bounded, we have

$$\begin{aligned}\left|\widehat{F}_T(\beta) - F_0(\beta)\right| &\leq \sup_{\beta\in\mathcal{B}}\left|\widehat{F}_T(\beta) - F_R(\beta)\right| + |F_R(\beta) - F_0(\beta)| \\ &= O_{\mathbb{P}}\left(R\sqrt{\frac{\log(RT)}{T}}\right) + o_R(1) \text{ a.e. } \beta\in\mathcal{B}.\end{aligned}$$

We can obtain an approximation order of $F_R(\beta) \underset{R\to\infty}{\longrightarrow} F_0(\beta)$ in R if we impose some degree of modulus on $f_0(\beta)$. We summarize the above result in the following theorem.

Theorem 5.3. *Suppose Assumptions in Theorem 5.2 hold. Further suppose $f_0(\beta)$ is uniformly bounded. Then,*

$$\left|\widehat{F}_T(\beta) - F_0(\beta)\right| = O_{\mathbb{P}}\left(R\sqrt{\frac{\log(RT)}{T}}\right) + o_R(1) \text{ a.e. } \beta\in\mathcal{B}.$$

where $o_R(1)\to 0$ as $R\to\infty$.

6 Fake Data Experiment

We conduct a fake data experiment in order to study the small sample properties of our estimator. We suppose that the true data generating process is indeed a random coefficients logit model. We assume that the econometrician observes aggregate market shares on T markets. Utility is defined by:

$$u_{ij} = x_j'\beta_i + \epsilon_{ij}$$

Thus, consumer i picks j when $x_j'\beta_i + \epsilon_{ij} > x_{j'}'\beta_i + \epsilon_{ij'}$ for $j'\neq j$. In our Monte Carlo, x_j is a 2×1 vector. We generate $x_{j,1}$ by making draws from a normal distribution with mean 1 and standard deviation 1, that is, $x_{j,1}\sim N(1,1)$. Also, $x_{j,2}$ has the distribution $N\left(0.8, 0.2^2\right) + 0.1x_{j,1}$.[19] There are $J=10$ products in each of our markets. J does not include the outside option. We will form a basis by using a mixture of normal distributions, as described in section 2. The number of basis points is $R=\frac{N}{5}$. The sample sizes N are 200, 500 and 1000. There is no measurement error; shares are exact.

Rather than a criterion such as integrated mean squared error, we prefer to test the structural use of our estimates. For each run, after we compute the estimate $\hat{f}(\beta)$, we evaluate its predictive power by drawing new product characteristics (from the same distributions) and predicting shares. We compare our results to those using the true $f^0(\beta)$. This is tests the structural use of discrete choice models to predict the demand for new goods.

[19] Independence between observable characteristics is not important for our estimator. Independence might improve estimates of $f(\beta)$.

We generate data using three alternative distributions of the random coefficients $f(\beta)$. In the first design, the tastes for characteristics are generated from a mixture of two normals:

$$0.7 \cdot N\left(\left[\begin{array}{c} 3 \\ 0 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) + 0.3 \cdot N\left(\left[\begin{array}{c} 1 \\ -1 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right).$$

All distributions of random coefficients will have a non-trivial variance matrix, while our basis functions are independent distributions of normal random coefficients. In the second design, the true coefficients are generated by a mixture of four normals:

$$\begin{aligned} &0.2 \cdot N\left(\left[\begin{array}{c} 3 \\ 0 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) + 0.4 \cdot N\left(\left[\begin{array}{c} 0 \\ 3 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) \\ &+0.3 \cdot N\left(\left[\begin{array}{c} 1 \\ -1 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right) + 0.1 \cdot N\left(\left[\begin{array}{c} -1 \\ 1 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right). \end{aligned}$$

In the third design and final design, the true coefficients are generated by the following mixture of six normals:

$$\begin{aligned} &0.1 \cdot N\left(\left[\begin{array}{c} 3 \\ 0 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) + 0.2 \cdot N\left(\left[\begin{array}{c} 0 \\ 3 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) \\ &+0.2 \cdot N\left(\left[\begin{array}{c} 1 \\ -1 \end{array}, \begin{array}{ccc} 0.1 & -0.1 \\ -0.1 & 0.1 \end{array}\right]\right) + 0.1 \cdot N\left(\left[\begin{array}{c} -1 \\ 1 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right) \\ &+0.3 \cdot N\left(\left[\begin{array}{c} 2 \\ 1 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right) + 0.1 \cdot N\left(\left[\begin{array}{c} 1 \\ 2 \end{array}, \begin{array}{ccc} 0.3 & 0.1 \\ 0.1 & 0.3 \end{array}\right]\right). \end{aligned}$$

Note that we estimate the for each of the designs only one; we do not perform a formal Monte Carlo study where we replicate our estimator with new datasets.

We summarize our fake data results in the table below. The first column states the distribution used to generate the random coefficients. The second column is the sample size N. The third column is the root mean squared error (RMSE) of our out of sample prediction for market shares. The RMSE is averaged over many new, counterfactual sets of products. It is not averaged over the statistical sampling error in the original estimator; we estimate for each design only one. The final column is the number of basis functions that have positive weight.

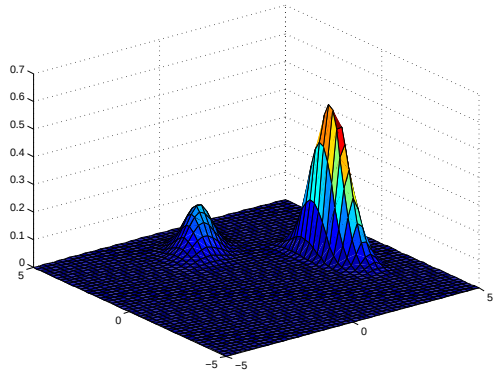
The results suggest that our estimator of $f(\beta)$ is excellent for predicting the market shares of new goods. To understand the units, with $J = 10$ and a mean share of 0.10, a prediction error of 0.01 (10% of 0.01) on each product corresponds to a RMSE of $\sqrt{10 \cdot 0.01^2/10} = 0.01$. Our results show the RMSE is between 0.02 and 0.01 when we have as few as 200 observations. When we have

1000 observations, the RMSE is typically around 0.001. Note that we are able to fit the observed market shares quite well with a fairly small number of basis functions. The number of nonzero basis functions ranges from 7 to 20 in the results reported below. Therefore, even when we make R large, most of these basis functions will have zero probability. This result is consistent with the literature on mixtures that demonstrates that even quite complicated distributed can be approximated with a surprisingly small number of mixture components.

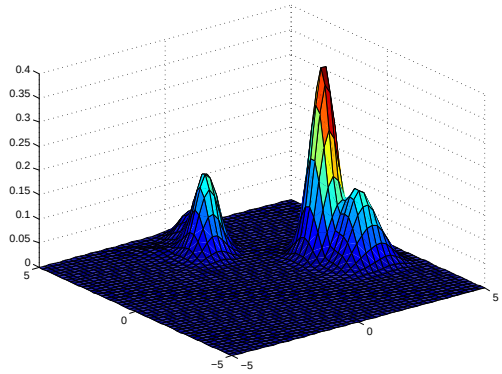
Table: Fake Data Results

Design	**Sample Size**	**RMSE**	**# of positive weights**
Normal Mixture 2	200	0.0201	7/40
	500	0.0031	11/100
	1000	0.0012	9/200
Normal Mixture 4	200	0.0279	8/40
	500	0.0015	19/100
	1000	0.0017	22/200
Normal Mixture 6	200	0.0138	7/40
	500	0.0013	21/100
	1000	0.0007	20/200

In the figures below, we plot the true versus the estimated distribution of $f(\beta)$. In all cases, the fitted distributions are based on our estimates when we have $N = 1000$ observations. A visual inspection of the densities shows that we are able to generate good fits to the true distribution of preferences using our estimator. Our Monte Carlo experiments demonstrate that it is possible to generate good estimates of market shares and the distribution of random coefficients with a modest number of observations.

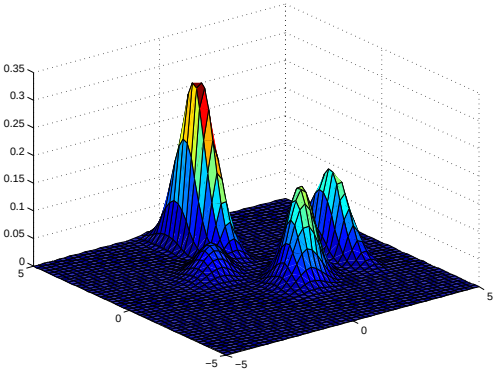


Truth

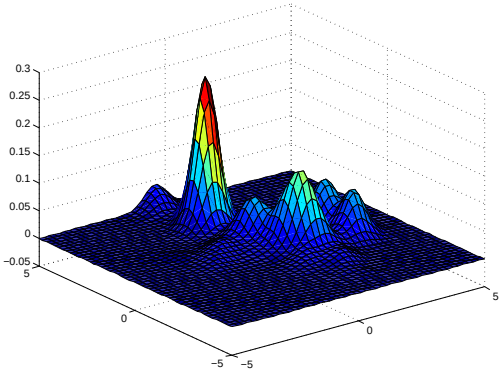


Estimated

Figure 1: **True and Fitted Distributions with Two Normal Mixtures**

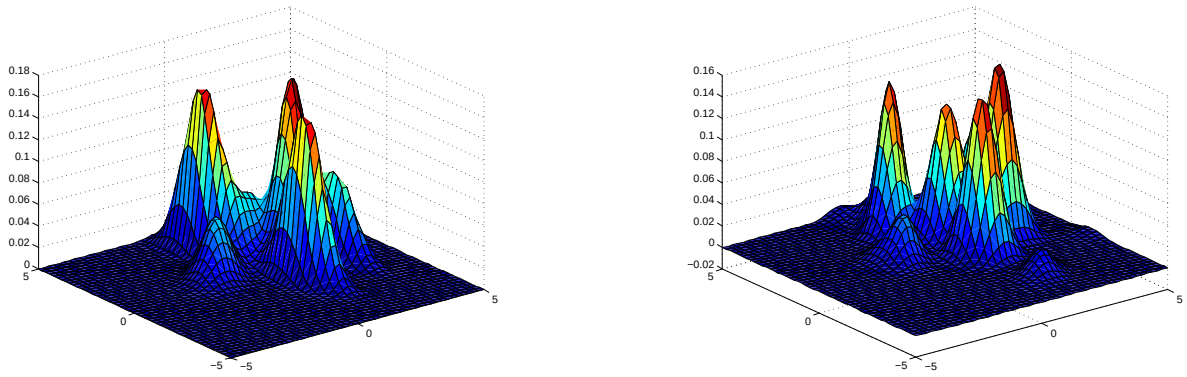


Truth



Estimated

Figure 2: **True and Fitted Distributions with Four Normal Mixtures**



Truth Estimated

Figure 3: **True and Fitted Distributions with Six Normal Mixtures**

7 Conclusion

In industrial organization, random coefficients are needed so that firms compete more closely with products that are similar in their characteristics. Researchers such as Berry, Levinsohn and Pakes (1995), Nevo (2001) and Petrin (2002) have found that including random coefficients in discrete choice models generates improved estimates of consumer demand, by which we mean demand elasticities that are a priori more sensible.

In this paper, we have proposed a new method for estimating the random coefficient logit model with a nonparametric distribution of random coefficients. Our method allows the researcher to drop standard parametric assumptions, such as independent normal random coefficients, that are commonly used in applied work. In terms of computer programming and execution time, our linear regression estimator is easier to work with than simulated likelihood or method of moments estimators for parametric models. Convergence of an optimization routine to the global optimum is much easier to show for linear regression with linear constraints than other statistical objective functions. Also, our estimator is much easier to program than nonparametric alternatives such as NPMLE using the EM algorithm, the Li and Barron (2000) iterative estimator, and Bayesian MCMC estimators.

Our estimator is useful for estimating dynamic programming discrete choice models. For dynamic programming, the estimator adds both a nonparametric distribution of random coefficients and actually cuts the computational time compared to the no-random coefficients model. The computational savings arise because we must solve the dynamic program only once for each basis vector. Adding an aggregate demand shock is important in practice. The BLP GMM estimator can be extended to our case. Overall, we wish to emphasize that random coefficients logit is just an example. Our estimator is a general mixtures estimator and can be applied to any economic model with unobserved heterogeneity.

A key question is identification: is only one distribution of random coefficients consistent with an ideal data set? Proving identification is model-specific, although we provide three general proof approaches: linear independence, bounded completeness, and identifying a distribution by its moments. We are the first to prove that the distribution of the random coefficients in the logit model is nonparametrically identified. Our logit identification result requires all characteristics to be continuous, but the characteristics only need to vary locally. We do not use identification at infinity.

Our other technical contribution focused on the asymptotic properties of the linear regression estimator. We derived the asymptotic distribution for the case where the true model has a finite set of types and found rates of convergence for the case where the true model may have an uncountable number of types. A concern for our linearly constrained estimators is that parameters may lie on the boundary (weights may have a density of 0), so that the distribution of the estimator is non-standard. However, to save on programming time, conservative OLS confidence regions can be used. We provide results for the location and scale model where the support of the approximation is estimated jointly with the basis weights. These same distribution results can be applied if parameters that are not random variables are estimated. For example, a researcher may want to include many product dummies, but estimating random coefficients on all the nuisance parameters may stretch the data too much.

We found in our Monte Carlo work that the method is able to generate good fits to consumer demand with reasonable sample sizes. Overall, we feel the bulk of the evidence shows that our estimator is easy to use for applied researchers. Random coefficient discrete choice models are used often in practice. We hope that by lowering the computational cost and raising the economic flexibility we will encourage researchers to use these methods even more.

A Regularity Conditions for Inequality Constrained Nonlinear Least Squares

We describe a set of regularity conditions under which we can justify consistency and asymptotic normality of ICNLS. Let $g(j,\cdot;\theta) = (g_1(j,\cdot;\theta)', g_2(j,\cdot;\theta)')'$ where $g_1(j,\cdot;\theta)$ and $g_1(j,\cdot;\theta)$ are defined in (25) and (26).

Assumption A.1. *θ_0 is the unique solution of (25) and (26).*

Assumption A.2. *(i) $\left\{\left(\widehat{P}(j,t), x'_{j,t}\right)', \ldots, \left(\widehat{P}(J,t), x'_{J,t}\right)'\right\}$ are independently distributed across $t = 1,\ldots,T$; (ii) For all $j = 1,\ldots,J$ and jointly, $g(j,\cdot;\theta) \in C^1$ (continuously differentiable w.r.t. θ) and satisfies the Lipschitz condition in $\theta \in \Theta$,*

$$\|g(j,\cdot;\theta_1) - g(j,\cdot;\theta_2)\| \leq B(\cdot)\|\theta_1 - \theta_2\| \text{ and } E[B(\cdot)^{2+\delta}] < \infty,\ \forall \theta_1,\theta_2 \in \Theta;$$

(iii) For all $j = 1,\ldots,J$ (jointly), $E\left[\sup_{\theta\in\Theta}\|g(j,\cdot;\theta)\|^{2+\delta}\right] < \infty$ for some $\delta > 0$; (iv) Θ is bounded.

Assumption A.2 implies uniform weak convergence (Andrews 1994),

$$\sup_{\theta\in\Theta}\left\|\frac{1}{T}\sum_t g(j,t;\theta) - E[g(j,t;\theta)]\right\| = o_{\mathbb{P}}(1). \tag{31}$$

Therefore, Assumption A.1 and Assumption A.2 imply that ICNLS is consistent

$$\widehat{\theta} \longrightarrow_p \theta_0.$$

Now we show the assumptions GMM2 to GMM5 of Andrews (2002) are satisfied for ICNLS under Assumption A.1 and A.2.

Assumption GMM2 of Andrews (2002) is satisfied as follows. GMM2 (a) holds due to the uniform convergence of (31). GMM2 (b) also holds because $\frac{1}{T}\sum_t g(j,t;\theta)$ is differentiable in θ. GMM2 (c) is satisfied by (25), (26), and Assumption A.1. GMM2 (d) is also satisfied because Assumption A.2 implies stochastic equicontinuity (Andrews 1994). Assumption GMM3 of Andrews (2002) holds because $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(g(1,t;\theta_0),\ldots,g(J,t;\theta_0))' \sim \mathcal{V} = \mathcal{N}(0,\pm)$ from the CLT by Assumption 2 (i) and (iii) where we define $\pm$ as

$$\begin{aligned}
\pm_{jj} &= E\left[\begin{pmatrix} h_1(j,\cdot;\theta_0)h_1(j,\cdot;\theta_0)' & h_1(j,\cdot;\theta_0)h_2(j,\cdot;\theta_0)' \\ h_2(j,\cdot;\theta_0)h_1(j,\cdot;\theta_0)' & h_2(j,\cdot;\theta_0)h_2(j,\cdot;\theta_0)' \end{pmatrix} e_{jt}^2\right] \text{ and} \\
\pm_{jj'} &= E\left[\begin{pmatrix} h_1(j,\cdot;\theta_0)h_1(j',\cdot;\theta_0)' & h_1(j,\cdot;\theta_0)h_2(j',\cdot;\theta_0)' \\ h_2(j,\cdot;\theta_0)h_1(j',\cdot;\theta_0)' & h_2(j,\cdot;\theta_0)h_2(j',\cdot;\theta_0)' \end{pmatrix} e_{jt}e_{j't}\right].
\end{aligned}$$

The assumption GMM4 of Andrews (2002) is satisfied because $I_b = (\frac{\partial I_b \cdot p}{\partial p'})$ has full row rank. Assumption GMM5 of Andrews (2002) is also satisfied because $\Lambda \equiv \Lambda_1 \times \Lambda_p$ is convex.

This proves the validity of the asymptotic distribution of $\widehat{\theta}$ given in (27).

B Asymptotic Distribution of the GMM estimator

Define $\Gamma = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \theta'} g(t;\theta)|_{\theta=\theta_0}$, $\mathcal{J}(A) = \Gamma' A \Gamma$, and $\mathcal{V} = \mathcal{N}(0,\Sigma)$ where

$$\Sigma = \text{plim} \frac{1}{T} \sum_{t=1}^{T} g(t;\theta_0) g(t;\theta_0)'.$$

The asymptotic distribution when f_0 is possibly on the boundary is

$$\sqrt{T}\left(\widehat{\theta}(A) - \theta_0\right) \underset{d}{\longrightarrow} \lambda^{***} \tag{32}$$

and $\lambda^{***} = \arg\inf_{\lambda \in \Lambda \equiv \Lambda_1 \times \Lambda_f} q(\lambda)$ where

$$\begin{aligned} q(\lambda) &= (\lambda - \mathcal{Z})' \mathcal{J}(A) (\lambda - \mathcal{Z}) \\ &\quad \mathcal{Z} \sim \mathcal{J}(A)^{-1} \Gamma' A \mathcal{V} \\ \Lambda_1 &= \left\{ \left(\lambda_1', \lambda_2'\right) \in \mathbb{R}^{2K} \right\}, \Lambda_f = \left\{ \lambda_f \in \mathbb{R}^R : \lambda_{f,1} + \ldots + \lambda_{f,R} = 0, I_b \lambda_f' \geq 0 \right\}, \end{aligned}$$

and I_b denotes the submatrix of an identity matrix I that consists of the rows of I such that $I_b f_0 = \mathbf{0}$.

We note that

$$\Gamma = \begin{bmatrix} H(1) \\ H(2) \\ \vdots \\ H(J-1) \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma(1,1) & \Sigma(1,2) & \cdots & \Sigma(1,J-1) \\ \Sigma(2,1) & \Sigma(2,2) & \cdots & \Sigma(2,J-1) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(J-1,1) & \Sigma(J-1,2) & \cdots & \Sigma(J-1,J-1) \end{bmatrix}$$

where

$$\begin{aligned} H(j) &= E \begin{bmatrix} h_1(j,\cdot;\theta_0) h_1(j,\cdot;\theta_0)' & h_1(j,\cdot;\theta_0) h_2(j,\cdot;\theta_0)' \\ h_2(j,\cdot;\theta_0) h_1(j,\cdot;\theta_0)' & h_2(j,\cdot;\theta_0) h_2(j,\cdot;\theta_0)' \end{bmatrix} \text{ and} \\ \Sigma(j,k) &= E \left[\begin{pmatrix} h_1(j,\cdot;\theta_0) h_1(k,\cdot;\theta_0)' e_j . e_k. & h_1(j,\cdot;\theta_0) h_2(k,\cdot;\theta_0)' e_j . e_k. \\ h_2(j,\cdot;\theta_0) h_1(k,\cdot;\theta_0)' e_j . e_k. & h_2(j,\cdot;\theta_0) h_2(k,\cdot;\theta_0)' e_j . e_k. \end{pmatrix} \right]. \end{aligned}$$

The asymptotic distribution of the optimal two step GMM estimator becomes

$$\sqrt{T}\left(\widehat{\theta}(\Sigma^{-1}) - \theta_0\right) \underset{d}{\longrightarrow} \lambda^{****} \tag{33}$$

and $\lambda^{****} = \arg\inf_{\lambda\in\Lambda\equiv\Lambda_1\times\Lambda_f} q(\lambda)$ where $q(\lambda) = (\lambda-\mathcal{Z})'\mathcal{J}(\Sigma^{-1})(\lambda-\mathcal{Z})$ and $\mathcal{Z} \sim \mathcal{J}(\Sigma^{-1})^{-1}\Gamma'\Sigma^{-1}\mathcal{V} = \mathcal{N}(0, \mathcal{J}(\Sigma^{-1})^{-1})$ noting that

$$\begin{aligned} Var[\mathcal{Z}] &= Var[\mathcal{J}(\Sigma^{-1})^{-1}\Gamma'\Sigma^{-1}\mathcal{V}] \\ &= \mathcal{J}(\Sigma^{-1})^{-1}\Gamma'\Sigma^{-1}Var[\mathcal{V}]\Sigma^{-1}\Gamma\mathcal{J}(\Sigma^{-1})^{-1} \\ &= \mathcal{J}(\Sigma^{-1})^{-1}\Gamma'\Sigma^{-1}\Gamma\mathcal{J}(\Sigma^{-1})^{-1} = \mathcal{J}(\Sigma^{-1})^{-1}. \end{aligned}$$

Now we verify a set of conditions under which the GMM estimator is consistent and follows the asymptotic distribution given by (32) or (33). First, note that Assumption A.2 for all $j = 1, \ldots, J$ (i.e., applying the uniform weak convergence of (31) for each j (and jointly) and the uniformly bounded $g(j, \cdot; \theta)$'s (Assumption A.2 (iii)) together with the triangular inequality) implies

$$\sup_{\theta\in\Theta} \left\| \left(\frac{1}{T}\sum_t g(t;\theta) \right)' A \left(\frac{1}{T}\sum_t g(t;\theta) \right) - (E[g(t;\theta)])' A (E[g(t;\theta)]) \right\| = o_{\mathbb{P}}(1). \tag{34}$$

Therefore, Assumption A.1 (identification) and Assumption A.2 implies the consistency of the GMM estimator such that

$$\widehat{\theta}(A) \xrightarrow[p]{} \theta_0.$$

Now we show the assumptions GMM2 to GMM5 of Andrews (2002) are satisfied for the GMM estimator.

Assumption GMM2 of Andrews (2002) is satisfied as follows. GMM2 (a) holds due to the uniform convergence of (31) for all $j = 1, \ldots, J$ and jointly. GMM2 (b) also holds since $\frac{1}{T}\sum_t g(j, t; \theta)$ is differentiable in θ for all $j = 1, \ldots, J$ and jointly. GMM2 (c) is satisfied by Assumption A.1 for all $j = 1, \ldots, J$ and jointly. GMM2 (d) is also satisfied since Assumption A.2 implies stochastic equicontinuity (Andrews 1994). For the optimal GMM estimator, assumption GMM2 (e) holds (applying the Slutsky theorem) since $\sup_{\theta\in\Theta:\|\theta-\theta_0\|=o(1)} \left\| \frac{1}{T}\sum_t g(t;\theta)g(t;\theta)' - \Sigma \right\|$ is implied by the uniform convergence of (31) for all $j = 1, \ldots, J$ (and jointly) and the continuity of $E\left[g(t;\theta)\, g(t;\theta)'\right]$ is implied by the dominated convergence theorem with the dominating function $\sup_{\theta\in\Theta} \|g(t;\theta)\|^2 < \infty$.

Assumption GMM3 of Andrews (2002) holds because $\frac{1}{\sqrt{T}}\sum_t g(t, m; \theta_0) \sim \mathcal{V} = \mathcal{N}(0, \Sigma)$ from the CLT, by Assumption A.2 (i) and (iii) for all $j = 1, \ldots, J$ and jointly. Assumption GMM4 of Andrews (2002) is satisfied because $I_b = (\frac{\partial I_b \cdot p}{\partial p'})$ has full row rank and not all of the $f_0(r)$'s are equal to zero at the same time. Assumption GMM5 of Andrews (2002) is satisfied since $\Lambda \equiv \Lambda_1 \times \Lambda_f$ is convex.

C Asymptotics for the Smoothed Mixture

Here we discuss the asymptotics for the smoothed mixing distribution with the continuous types, proposed in Section 14. We use the following equation,

$$\widehat{P}(j,t) = \sum_{r=1}^{R} \bar{f}(r) \int \left(\frac{\exp\left(x'_{j,t}\beta\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta\right)} \right) h(\beta, r) d\beta + e_{jt} \tag{35}$$

where $h(\beta, r) = N(\beta|\mu^{(r)}, \sigma^{(r)})$ and the final approximation is $f(\beta) = \sum_{r=1}^{R} \bar{f}(r) h(\beta, r)$.

Now suppose we approximate the expectation of $\left(\frac{\exp(x'_{j,t}\beta)}{\sum_{j'=1}^{J} \exp(x'_{j',t}\beta)} \right)$ taken over a particular normal distribution whose density is $h(\beta, r)$ by a finite number of points, as in

$$\int \frac{\exp\left(x'_{j,t}\beta\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta\right)} h(\beta, r) d\beta \approx \frac{1}{S} \sum_{s=1}^{S} \frac{\exp\left(x'_{j,t}\beta_{(s)}^{(r)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(s)}^{(r)}\right)}$$

where the $\left\{\beta_{(s)}^{(r)}\right\}_{s=1}^{S}$ are drawn from $h(\beta, r)$. This implies (35) can be approximated as

$$\begin{aligned} \widehat{P}(j,t) &= \sum_{r=1}^{R} \bar{f}(r) \left\{ \frac{1}{S} \sum_{s=1}^{S} \frac{\exp\left(x'_{j,t}\beta_{(s)}^{(r)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(s)}^{(r)}\right)} \right\} + e_{jt} \\ &= \frac{f(1)}{S} \frac{\exp\left(x'_{j,t}\beta_{(1)}^{(1)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(1)}^{(1)}\right)} + \ldots + \frac{f(1)}{S} \frac{\exp\left(x'_{j,t}\beta_{(S)}^{(1)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(S)}^{(1)}\right)} + \ldots \\ &\quad + \frac{f(R)}{S} \frac{\exp\left(x'_{j,t}\beta_{(1)}^{(R)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(1)}^{(R)}\right)} + \ldots + \frac{f(R)}{S} \frac{\exp\left(x'_{j,t}\beta_{(S)}^{(R)}\right)}{\sum_{j'=1}^{J} \exp\left(x'_{j',t}\beta_{(S)}^{(R)}\right)} + e_{jt}. \end{aligned} \tag{36}$$

Therefore, we can interpret (36) as the model (29) with $R \cdot S$ regressors and $R \cdot (S-1)$ restrictions where the first group of S coefficients on the first group of S regressors are identical to $\frac{p(1)}{S}$, the second group of S coefficients on the second group of S regressors are identical to $\frac{p(2)}{S}$, and so on. Therefore, the smoothed mixture model can be nested in the discrete mixture model.

D Alternative Estimators that Adapt to the Sparsity of Types

Consider a model with a finite number of types. Until now we have not considered the fact that the number of types (R) generated picked by the researcher, i.e., the number of basis functions,

may exceed the number of true types, denoted by

$$R\left(f_0\right) = \sum_{r=1}^{R} I\left\{f_0(r) > 0\right\}$$

where $f_0(r)$ is the true mass of type r. If a generated basis component does not belong to the set of true basis functions, we expect the estimated weight on such a basis will be close to zero. Consequently, we may have many coefficients close to zero. For this sort of problem, the statistical literature proposes alternative estimators that adaptively estimate such zeros and non-zeros. Estimators holding such a property have been often called *oracle* estimators.

We may estimate $f(\cdot)$ using penalized inequality constrained least squares (PICLS) such that

$$\widehat{f}^* = \arg\min_{f \in \Delta^R} \left\{ \sum_{j=1}^{J-1} \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{P}(j, x_t) - d_f(j, x_t) \right)^2 + \text{pen}\left(f\right) \right\} \tag{37}$$

where pen(f) is a penalizing function. Bunea, Tsybakov, and Wegkamp (2006)'s proposal for a penalty function is

$$\text{pen}\left(f\right) = 2 \sum_{j=1}^{J-1} \sum_{r=1}^{R} \xi_T(R) \left\| d(r, \cdot) \right\|_T \left| f(r) \right|$$

where $\left\| d(r,\cdot) \right\|_T = \sqrt{\frac{1}{T} \sum_{t=1}^{T} d\left(r, j, x_t\right)^2}$, $\xi_T(R) = C\sqrt{\frac{\log(RT)}{T}}$, and $\Delta^R = \left\{ f \in \mathbb{R}^R : f(r) \geq 0 \,\text{and}\, \sum_{r=1}^{R} f(r) = 1 \right\}$. Now we strengthen Assumption 5.3 as follows.

Assumption D.1. *There exist $f^* \in \Delta^R$ and a constant $C_* > 0$ such that $\frac{1}{J-1} \sum_{j=1}^{J-1} ||d_{f^*}(j,x) - P(j,x,f_0)||_\infty^2 \leq C_* \eta_T^2(R) R(f^*)$ for a sequence $\eta_T(R) \underset{T\to\infty}{\longrightarrow} 0$.*

Assumption D.1 means that the unknown function $P\left(j, x, f_0\right)$ can be well approximated by quite smaller subsets of basis functions only. This assumption holds if $P\left(j, x, f_0\right)$ belongs to a parametric family. In a nonparametric context, this assumption means that $P\left(j, x, f_0\right)$ is well approximated by the linear span of $R\left(f^*\right)$ (possibly $\ll R$) basis functions only. Following Bunea, Tsybakov, and Wegkamp (2006), one can show that

Theorem D.1. *Suppose Assumptions 5.1, 5.2 and D.1 hold. Further suppose $\eta_T = \sqrt{A\frac{\log(RT)}{T}}$ with a suitable large constant A and $\frac{R^2 \log R}{T} \to 0$. Then, we have*

$$\begin{aligned} \sum_{j=1}^{J-1} \left\| d_{\widehat{f}^*}\left(j, x_t\right) - P\left(j, x_t, f_0\right) \right\|_{MT}^2 &= O_{\mathbb{P}}\left(\frac{J \cdot R\left(f^*\right) \cdot \log(RT)}{T} \right) \text{ and} \\ \sum_{r=1}^{R} \left| \widehat{f}^*\left(r\right) - f^*\left(r\right) \right| &= O_{\mathbb{P}}\left(R\left(f^*\right) \sqrt{\frac{\log(RT)}{T}} \right). \end{aligned}$$

where $R(f) = \sum_{r=1}^{R} I\{f(r) > 0\}$.

Therefore, we have a sharper bound compared to the original estimator $\widehat{f}$ in (30) when $R \gg R(f^*)$. Note that the implication of Theorem D.1 is quite strong in that the estimator $\widehat{f}^*$ adapts to the possible sparsity of the types even when we do not know the exact type distribution. The convergence rate is faster when there are less consumer types ($R(f^*) \ll R$). Moreover, if we indeed knew the type distribution for a parametric case $f_0 = f^*$ (in other words, we knew which of true $f^*(r)$'s are positive and others are zero), we would obtain $\sum_{r=1}^{R} \left|\widehat{f}^*(r) - f^*(r)\right| = O_p\left(\sqrt{\frac{1}{T}} R(f^*)\right)$, which is the rate of the OLS. The result of Theorem D.1 suggests that the loss in the approximation precision due to not having prior knowledge of the type distribution is quite negligible. It is just an order of the root of logarithm.

Another type of penalty function is proposed by Huang, Horowitz, and Ma (2007) (HHM). We can change $\text{pen}(f)$ in (37) with

$$\text{pen}(f) = \lambda_n \sum_{r=1}^{R} |f(r)|^{\gamma} \text{ with } 0 < \gamma < 1$$

according to their proposal. We call the resulting estimator, denoted by $\widehat{f}^{**}$, the bridge estimator:

$$\widehat{f}^{**} = \text{argmin}_{f \in \Delta^R} \left\{ \sum_{j=1}^{J-1} \frac{1}{T} \sum_{t=1}^{T} \left\{ \widehat{P}(j, x_t) - d_f(j, x_t) \right\}^2 + \lambda_n \sum_{r=1}^{R} |f(r)|^{\gamma} \right\}.$$

HHM show that this bridge estimator can correctly distinguish between nonzero and zero coefficients and the estimator of the nonzero coefficients has the same asymptotic distribution as if we knew the nonzero terms with certainty.

However, these gains do not come without a risk. Leeb and Pötscher (2005) warn that any estimator that shares the oracle property may not be robust to some local alternatives. As a result, the maximal risk (e.g. maximal mean squared error of $\widehat{f}^*$ or $\widehat{f}^{**}$ taken over the whole parameter space uniformly) of this sort of estimator can diverge to infinity. Therefore, these oracle estimators could perform worse than our simple estimator $\widehat{f}$ in some cases.

E Mathematical Proofs

Lemma E.1. *Suppose Assumption 5.1 and 5.2 hold. Then*

$$\min\left[\Pr\left\{\|d(r,\cdot)\|_T^2 \leq 2\|d(r,\cdot)\|_0^2,\ \forall r\right\}, \Pr\left\{\|d(r,\cdot)\|_0 \leq 2\|d(r,\cdot)\|_T,\ \forall r\right\}\right] \geq 1 - R\exp\left(-\frac{Tc_0^2}{2\zeta_0^2}\right)$$

Proof. Note that $\|d(r,\cdot)\|_0 \geq c_0 > 0$ for all $r = 1, \ldots, R$ unless $\mathcal{X}$ has no positive mass. Then the claim follows immediately from the union bound and Hoeffding's inequality. □

Lemma E.2. *Let $\mathcal{D}$ be the linear space spanned by some functions $d(1,\cdot), \ldots, d(q,\cdot)$ such that*

$$d(r,\cdot) \in \mathcal{D}_0 = \left\{ d : d = \left(\exp\left(x_j'\beta\right) / \sum_{j'=1}^{J} \exp\left(x_{j'}'\beta\right) \right), x \in \mathcal{X}, \beta \in \mathcal{B} \right\}.$$

Then

$$\Pr\left\{ \sup_{d \in \mathcal{D}\setminus\{0\}} \frac{\|d\|_0^2}{\|d\|_T^2} > 2 \right\} \leq q^2 \exp\left(-C \frac{T}{\zeta_0^4 q^2} \right) \tag{38}$$

for some constant C.

Proof. Let $\phi_1, \ldots, \phi_N$ be an orthonormal basis of $\mathcal{D}$ in L_2 with $N \leq q$. Also let $\overline{\rho}(A)$ denote the following quantity for a symmetric matrix A:

$$\overline{\rho}(A) = \sup_{\lambda,\lambda'} \sum_{\lambda} |a_\lambda| \sum_{\lambda'} |a_{\lambda'}| \left| A_{\lambda,\lambda'} \right|,$$

where the sup is taken over sequences $\{a_\lambda\}_{\lambda=1}^N$ with $\sum_\lambda a_\lambda^2 = 1$. Then, from Lemma 5.2 in Baraud (2002),

$$\Pr\left\{ \sup_{d \in \mathcal{D}\setminus\{0\}} \frac{\|d\|_0^2}{\|d\|_T^2} > c \right\} \leq N^2 \exp\left(-T \frac{(\mu_0 - c^{-1})^2}{4\mu_1 \max\{\overline{\rho}^2(A), \overline{\rho}(B))\}} \right) \tag{39}$$

where $A_{\lambda,\lambda'} = \sqrt{E[\phi_\lambda^2 \phi_{\lambda'}^2]}$ and $B_{\lambda,\lambda'} = \|\phi_\lambda \phi_{\lambda'}\|_\infty$ for $\lambda, \lambda' = 1, \ldots, N$ and μ_0 and μ_1 denote the lower bound and upper bound of the density of X, respectively. We find $\left|A_{\lambda,\lambda'}\right| \leq \zeta_0^2$ and $\left|B_{\lambda,\lambda'}\right| \leq \zeta_0^2$. It follows that

$$\overline{\rho}(A) \leq \zeta_0^2 \sup_{\lambda,\lambda'} \sum_{\lambda} |a_\lambda| \sum_{\lambda'} |a_{\lambda'}| = \zeta_0^2 \sup_{\lambda} \left(\sum_{\lambda} |a_\lambda| \right)^2 \leq \zeta_0^2 \sup_{\lambda} N \sum_{\lambda} |a_\lambda|^2 = \zeta_0^2 N \leq \zeta_0^2 q$$

where $\left(\sum_\lambda |a_\lambda|\right)^2 \leq N \sum_\lambda |a_\lambda|^2$ holds by the Cauchy-Schwarz inequality. Similarly we have $\overline{\rho}(B) \leq \zeta_0^2 q$. Noting $N \leq q$, from (39), we conclude

$$\Pr\left\{ \sup_{d \in \mathcal{D}\setminus\{0\}} \frac{\|d\|_0^2}{\|d\|_T^2} > 2 \right\} \leq q^2 \exp\left(-C \frac{T}{\zeta_0^4 q^2} \right)$$

with $C = \frac{(\mu_0 - 1/2)^2}{4\mu_1}$. □

Now we define

$$\Psi_{T,R} = \left(\frac{1}{T} \sum_{t=1}^{T} d\left(r, j, x_t\right) d\left(r', j, x_t\right) \right)_{1 \leq r, r' \leq R} \text{ and } \Psi_R = \left(E\left[d\left(r, j, x_t\right) d\left(r', j, x_t\right) \right] \right)_{1 \leq r, r' \leq R},$$

where we suppress $\Psi_{T,R}$ and Ψ_R's dependence on j for notational simplicity.

Lemma E.3. *Suppose Assumption 5.1 and 5.2 hold. Then,*

$$\Pr\left\{\Psi_{T,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2} diag\left(\Psi_{T,R}\right) \geq 0\right\} \geq 1 - R\exp\left(-\frac{Tc_0^2}{2\zeta_0^2}\right) - R^2\exp\left(-C\frac{T}{\zeta_0^4 R^2}\right).$$

Proof. First, note that $\Psi_R - \frac{\xi_{\min}(R)}{\zeta_0^2}\text{diag}\left(\Psi_R\right) \geq 0$ (positive semi-definite) since Ψ_R is a positive definite matrix by Assumption 5.2 (ii). Now let $\mathcal{D}$ be the linear space spanned by $d(1,\cdot),\ldots,d(R,\cdot)$. Now note that under the event $\|d(r,\cdot)\|_T^2 \leq 2\|d(r,\cdot)\|_0^2\ \forall r = 1,\ldots,R$, we have $\text{diag}\left(\Psi_{T,R}\right) \leq 2\text{diag}\left(\Psi_R\right)$ and note that under the event $\left\{\sup_{d\in\mathcal{D}\setminus\{0\}}\frac{\|d\|_0^2}{\|d\|_T^2} > 2\right\}$, we have $\Psi_{T,R} \geq \Psi_R/2$. Therefore, under the event $\|d(r,\cdot)\|_T^2 \leq 2\|d(r,\cdot)\|_0^2,\ \forall r = 1,\ldots,R \cap \left\{\sup_{d\in\mathcal{D}\setminus\{0\}}\frac{\|d\|_0^2}{\|d\|_T^2} > 2\right\}$, we have

$$\Psi_{T,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2}\text{diag}\left(\Psi_{T,R}\right) \geq \Psi_R/2 - \frac{\xi_{\min}(R)}{4\zeta_0^2}2\text{diag}\left(\Psi_R\right) \geq 0$$

since $\Psi_R - \frac{\xi_{\min}(R)}{\zeta_0^2}\text{diag}\left(\Psi_R\right) \geq 0$. It follows that

$$\begin{aligned}
&1 - \Pr\left\{\Psi_{T,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2}\text{diag}\left(\Psi_{T,R}\right) \geq 0\right\} \leq \\
&1 - \Pr\left\{\|d(r,\cdot)\|_T^2 \leq 2\|d(r,\cdot)\|_0^2,\ \forall r = 1,\ldots,R\right\} + 1 - \Pr\left\{\sup_{d\in\mathcal{D}\setminus\{0\}}\frac{\|d\|_0^2}{\|d\|_T^2} > 2\right\} \\
\leq\ & R\cdot\exp\left(-\frac{Tc_0^2}{2\zeta_0^2}\right) + R^2\cdot\exp\left(-C\frac{T}{\zeta_0^4 R^2}\right)
\end{aligned}$$

where the last inequality is obtained from Lemma E.1 and Lemma 38. □

Lemma E.4. *Suppose Assumptions 5.1 and 5.2 hold. Then,*

$$\begin{aligned}
&\Pr\left\{\left|\frac{1}{T}\sum_{t=1}^{T} e_{jt}d(r,j,x_t)\right| \leq \eta_T\|d(r,j,x_t)\|_T \ \textit{ for all } r = 1,\ldots,R\right\} \qquad (40)\\
\geq\ & 1 - 2R\cdot\exp\left(-\frac{T\eta_T^2(R)}{8}\right) - 2R\cdot\exp\left(-\frac{T\eta_T(R)c_0}{8\zeta_0}\right) - R\cdot\exp\left(-\frac{Tc_0^2}{2\zeta_0^2}\right).
\end{aligned}$$

Proof. We will apply a version of Bernstein's inequality (Hoeffding 1963; Yurinskii 1976): Let $Z_1,\ldots,Z_n$ be independent random variables such that i) $E\left[Z_i\right] = 0$ ii) $E\left[Z_i^2\right] = b_i$ iii) $E\left[|Z_i|^q\right] \leq \frac{b_i}{2}c^{q-2}q!$ for some positive constant c independent of i and for all $q \geq 2$. Then for any $\varepsilon > 0$, we

have

$$\Pr\left\{\left|\sum_{i=1}^{n} Z_i\right| \geq n\varepsilon\right\} \leq 2\exp\left(-\frac{n^2\varepsilon^2}{2\left(n\max_{i\leq n}\{b_i\} + cn\varepsilon\right)}\right) \tag{41}$$

noting $n\max_{i\leq n}\{b_i\} \geq \sum_{i=1}^{n} b_i$. By Assumptions 5.1 and 5.2, we have

$$\begin{aligned}
&\frac{1}{T}\sum_t E\left[\left|e_{jt}d\left(r,j,X_t\right)\right|^q|X_1,\ldots,X_T\right] \\
\leq\ & \zeta_0^{q-2}\frac{1}{T}\sum_t \left|d\left(r,j,X_t\right)\right|^2 E\left[|e_{jt}|^q|X_1,\ldots,X_T\right] \\
\leq\ & q!\zeta_0^{q-2}\left\|d(r,\cdot)\right\|_T^2 = \frac{\left(\left\|d(r,\cdot)\right\|_T\sqrt{2}\right)^2}{2}\zeta_0^{q-2}q!
\end{aligned}$$

since $E\left[|e_{jt}|^q|X_1,\ldots,X_T\right] \leq 1$ by construction of $|e_{jt}| = \left|\widehat{P}\left(j,t\right) - P\left(j,t\right)\right| \leq 1$ uniformly.

Noting $\|d(r,\cdot)\|_0 > 0$ for all r as long as $\mathcal{X}$ has some positive mass, using (41) and the union bound, we obtain

$$\begin{aligned}
& E\left[\Pr\left\{\left|\frac{1}{T}\sum_t e_{jt}d(r,j,X_t)\right| \geq \eta_T\left\|d(r,\cdot)\right\|_T,\ \forall r=1,\ldots,R\middle| X_1,\ldots,X_T\right\}\right] \qquad (42)\\
\leq\ & E\left[2\sum_{r=1}^{R}\exp\left(-\frac{T\eta_T^2\|d(r,\cdot)\|_T^2}{2\left(\|d(r,\cdot)\|_T^2 2 + \zeta_0\eta_T\left\|d(r,\cdot)\right\|_T\right)}\right)\right] \\
\leq\ & E\left[2R\exp\left(-\frac{T\eta_T^2\left\|d(r,\cdot)\right\|_T^2}{4\left\|d(r,\cdot)\right\|_T^2 2}\right)\right] + 2\sum_{r=1}^{R}E\left[\exp\left(-\frac{T\eta_T^2\|d(r,\cdot)\|_T^2}{4\zeta_0\eta_T\left\|d(r,\cdot)\right\|_T}\right)\right] \\
=\ & 2R\exp\left(-\frac{T\eta_T^2}{8}\right) + 2\sum_{r=1}^{R}E\left[\exp\left(-\frac{T\eta_T\left\|d(r,\cdot)\right\|_T}{4\zeta_0}\right)\right]
\end{aligned}$$

where the last inequality is obtained from the fact that

$$\exp\left(-x/2a\right) + \exp\left(-x/2b\right) \geq \exp\left(-x/\left(a+b\right)\right)$$

for $x, a, b > 0$. Finally note under the event $\{\|d(r,\cdot)\|_0 \leq 2\left\|d(r,\cdot)\right\|_T,\ \forall r=1,\ldots,R\}\}$,

$$\begin{aligned}
2\sum_{r=1}^{R}\exp\left(-\frac{T\eta_T\left\|d(r,\cdot)\right\|_T}{4\zeta_0}\right) &\leq 2\sum_{r=1}^{R}\exp\left(-\frac{T\eta_T\|d(r,\cdot)\|/2}{4\zeta_0}\right) \qquad (43)\\
&\leq 2\sum_{r=1}^{R}\exp\left(-\frac{T\eta_T c_0/2}{4\zeta_0}\right) = 2R\exp\left(-\frac{T\eta_T c_0}{8\zeta_0}\right).
\end{aligned}$$

From (42) and (43), we obtain the result since $\Pr\{\|d(r,\cdot)\|_0 > 2\|d(r,\cdot)\|_T,\ \forall r = 1,\dots,R\} = R\exp(-\frac{Tc_0^2}{2\zeta_0^2})$ by Lemma E.1. □

Lemma E.5. *Suppose Assumptions 5.1 and 5.2 hold. Then, for any $T \geq 1$, $R \geq 2$, and $a > 1$, we have*

$$\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)\right\|_T^2 \leq \frac{a+1}{a-1}\sum_{j=1}^{J-1}\|d_f(j,x_m) - P(j,x_t,f_0)\|_T^2 + \frac{\zeta_0^2}{\xi_{\min}(R)}\frac{8a^2}{a-1}\eta_T^2 R\cdot(J-1)$$

for all $f \in \Delta^R$ with probability $\geq 1 - p_{T,R}$ where

$$p_{T,R} = 2RJ\exp\left(-\frac{T\eta_T^2}{8}\right) + 2RJ\exp\left(-\frac{T\eta_T c_0}{8\zeta_0}\right) + 2RJ\exp\left(-\frac{Tc_0^2}{2\zeta_0^2}\right) + R^2J\exp\left(-C\frac{T}{\zeta_0^2R^2}\right)$$

Proof. By construction of $d_{\widehat{f}}(\cdot)$, we can write

$$\begin{aligned}
&\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)\right\|_T^2 \\
\leq\ &\sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2 + \sum_{j=1}^{J-1}\frac{2}{T}\sum_t e_{jt}(d_{\widehat{f}}(j,x_t) - d_f(j,x_t)).
\end{aligned}$$

Now let $V_{T,r}(j) = \frac{1}{T}\sum_{t=1}^T e_{jt}d(r,j,x_t)$ and recall the $R\times R$ matrix $\Psi_{T,R}$. Then, we have

$$\frac{1}{T}\sum_{t=1}^T e_{jt}(d_{\widehat{f}} - d_f)(j,x_t) = \sum_{r=1}^R V_{T,r}(j)(\widehat{f}(r) - f(r))$$

and we obtain by the triangle inequality,

$$\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)\right\|_T^2 \leq \sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2 + 2\sum_{j=1}^{J-1}\sum_{r=1}^R |V_{T,r}(j)|\left|\widehat{f}(r) - f(r)\right|. \tag{44}$$

Define two events $E_{0,j} = \cap_{r=1}^R\{|V_{T,r}(j)| \leq \eta_T\|d(r,\cdot)\|_T\}$ for some positive sequence η_T and $E_{1,j} =$

$\{\Psi_{T,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2}\text{diag}(\Psi_{T,R}) \geq 0\}$. Then, under $E_{0,j} \cap E_{1,j}$, we note

$$\begin{aligned}
\sum_{r=1}^{R} V_{T,r}^2(j) \left|\widehat{f}(r) - f(r)\right|^2 &\leq \eta_T^2 \sum_{r=1}^{R} \|d(r,\cdot)\|_T^2 \left|\widehat{f}(r) - f(r)\right|^2 \qquad (45)\\
&= \eta_T^2 \frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R}\left(\widehat{f}(r) - f(r)\right)^2 d(r,j,x_t)^2\\
&= \eta_T^2 (\widehat{f} - f)' \text{diag}(\Psi_{T,R}(j))(\widehat{f} - f)\\
&\leq \eta_T^2 \left(\frac{\xi_{\min}(R)}{4\zeta_0^2}\right)^{-1} (\widehat{f} - f)'\Psi_{T,R}(j)(\widehat{f} - f)\\
&= \eta_T^2 \left(\frac{\xi_{\min}(R)}{4\zeta_0^2}\right)^{-1} \left\|d_{\widehat{f}}(j,x_t) - d_f(j,x_t)\right\|_T^2.
\end{aligned}$$

From (44) and (45), it follows that on $\cap_{j=1}^{J-1}(E_{0,j} \cap E_{1,j})$

$$\begin{aligned}
&\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)\right\|_T^2\\
\leq\ &\sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2 + 2\sum_{j=1}^{J-1}\sum\nolimits_{r=1}^{R}|V_{T,r}(j)|\left|\widehat{f}(r) - f(r)\right|\\
\leq\ &\sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2 + 2\sum_{j=1}^{J-1}\sqrt{R}\sqrt{\sum\nolimits_{r=1}^{R} V_{T,r}^2(j)\left|\widehat{f}(r) - f(r)\right|^2}\\
\leq\ &\sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2 + 2\sum_{j=1}^{J-1}\eta_T\sqrt{R\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1}}\left\|d_{\widehat{f}}(j,x_t) - d_f(j,x_t)\right\|_T\\
\leq\ &\sum_{j=1}^{J-1}\|d_f(j,x_t) - P(j,x_t,f_0)\|_T^2\\
&+2\sum_{j=1}^{J-1}\eta_T\sqrt{R\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1}}\left(\left\|d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)\right\|_T + \|d_f(j,x_t) - P(j,x_t,f_0)\|_T\right)
\end{aligned}$$

by the Cauchy-Schwarz inequality and the triangle inequality. Applying the inequality $2xy \leq x^2a + y^2/a$ (any x,y, $a > 0$) to $x = \eta_T\sqrt{R\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1}}$ and $y = ||d_{\widehat{f}}(j,x_t) - P(j,x_t,f_0)||_T$ and to $x = \eta_T\sqrt{R\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1}}$ and $y = \|d_f(j,x_t) - P(j,x_t,f_0)\|_T$, respectively, we obtain under

$\prod_{j=1}^{J-1} E_{0,j} \cap E_{1,j}$,

$$\begin{aligned}
&\sum_{j=1}^{J-1} \left\| d_{\widehat{f}}(j, x_t) - P(j, x_t, f_0) \right\|_T^2 \\
\leq\ & \sum_{j=1}^{J-1} \| d_f(j, x_t) - P(j, x_t, f_0) \|_T^2 + \sum_{j=1}^{J} \left\| d_{\widehat{f}}(j, x_t) - P(j, x_t, f_0) \right\|_T^2 / a \\
&+ \sum_{j=1}^{J-1} \| d_f(j, x_t) - P(j, x_t, f_0) \|_T^2 / a + 2a\eta_T^2 R \left(\xi_{\min}(R)/4\zeta_0^2 \right)^{-1} \cdot (J-1)
\end{aligned}$$

It follows that under $\bigcap_{j=1}^{J-1} (E_{0,j} \cap E_{1,j})$, for all $a > 1$,

$$\sum_{j=1}^{J-1} \left\| d_{\widehat{f}}(j, x_t) - P(j, x_t, f_0) \right\|_T^2 \leq \frac{a+1}{a-1} \sum_{j=1}^{J-1} \| d_f(j, x_t) - P(j, x_t, f_0) \|_T^2 + \frac{\zeta_0^2}{\xi_{\min}(R)} \frac{8a^2}{a-1} \eta_T^2 R (J-1).$$

The conclusion follows from

$$\Pr\left\{ \left(\bigcap_{j=1}^{J-1} E_{0,j} \right)^C \right\} \leq \sum_{j=1}^{J-1} \Pr\left\{ E_{0,j}^C \right\} = (J-1)R \left\{ 2\exp\left(-\frac{T\eta_T^2}{8} \right) + 2\exp\left(-\frac{T\eta_T c_0}{8\zeta_0} \right) + \exp\left(-\frac{Tc_0^2}{2\zeta_0^2} \right) \right\}$$

by Lemma 40 and $\Pr\left\{ \left(\bigcap_{j=1}^{J-1} E_{1,j} \right)^C \right\} \leq \sum_{j=1}^{J-1} \Pr\left\{ E_{1,j}^C \right\} = (J-1) \left\{ R\exp\left(-\frac{Tc_0^2}{2\zeta_0^2} \right) + R^2 \exp\left(-C\frac{T}{16\zeta_0^4 R^2} \right) \right\}$ by Lemma E.3. □

E.1 Proof of Theorem 5.1

From Lemma E.5, we find that the fastest convergence rate will be obtained when the order of η_T is as small as possible while keeping the convergence of $p_{T,R}$ to zero. By inspecting $p_{T,R} = (J-1) \left\{ 2R\exp\left(-\frac{T\eta_T^2}{8} \right) + 2R\exp\left(-\frac{T\eta_T c_0}{8\zeta_0} \right) + 2R\exp\left(-\frac{Tc_0^2}{2\zeta_0^2} \right) + R^2 \exp\left(-C\frac{T}{\zeta_0^2 R^2} \right) \right\}$, we note that the optimal rate is obtained when $\eta_T(R) = \sqrt{A\frac{\log(RT)}{T}}$ with a constant $A > 8/3$ since the first term in $p_{T,R}$ dominates the second term in $p_{T,R}$ when η_T is small and $p_{T,R} \to 0$ with $\eta_T(R) = \sqrt{A\frac{\log(RT)}{T}}$. The inspection of the third and the fourth term in $p_{T,R}$ reveals that we also require R should satisfy $\frac{R^2 \log(R)}{T} \to 0$ so that $p_{T,R} \to 0$.

The result of Theorem 5.1 is immediately obtained from these requirements and Lemma E.5

such that

$$\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t)-P(j,x_t,f_0)\right\|_T^2=O_{\mathbb{P}}\left(\max\left\{\frac{J\cdot R\log(RT)}{T},\sum_{j=1}^{J-1}\|d_f(j,x_t)-P(j,x_t,f_0)\|\right\}\right)$$

since $\frac{\eta_0^2}{\xi_{\min}(R)}\frac{8a^2}{a-1}\eta_T^2R=O\left(\frac{\log(RT)}{T}R\right)$ under $\eta_T=\sqrt{A\frac{\log(RT)}{T}}$ and Lemma E.5 holds for any $f\in\Delta^R$.

E.2 Proof of Theorem 5.2

The first result of Theorem 5.2 immediately follows from Theorem 5.1 combined with Assumption 5.3. Now we show the second claim. Note that with probability approaching to one (by Lemma E.1 and $\|d(r,j,x)\|_0\geq c_0>0$ for all $r=1,\ldots,R$ and $j=1,\ldots,J$ unless $\mathcal{X}$ has no mass), we have

$$\begin{aligned}
&\frac{c_0}{2}\sum_{r=1}^{R}\left|\widehat{f}(r)-f^*(r)\right| \\
\leq\ & \frac{1}{J-1}\sum_{j=1}^{J-1}\sum_{r=1}^{R}\|d(r,j,x_t)\|_T\left|\widehat{f}(r)-f^*(r)\right| \leq \frac{1}{J-1}\sum_{j=1}^{J-1}\sqrt{R}\sqrt{\sum_{r=1}^{R}\|d(r,j,x_t)\|_T^2\left|\widehat{f}(r)-f^*(r)\right|^2} \\
\leq\ & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}}\sqrt{R}\frac{1}{J-1}\sum_{j=1}^{J-1}\left\|d_{\widehat{f}}(j,x_t)-d_{f^*}(j,x_t)\right\|_T \\
\leq\ & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}}\sqrt{R}\frac{1}{J-1}\sum_{j=1}^{J-1}\left(\left\|d_{\widehat{f}}(j,x_t)-P(j,x_t,f_0)\right\|_T+\|d_{f^*}(j,x_t)-P(j,x_t,f_0)\|_T\right) \\
\leq\ & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}}\sqrt{R}\left(\sqrt{\frac{1}{J-1}\sum_{j=1}^{J-1}||d_{\widehat{f}}(j,\cdot)-P(j,\cdot,f_0)||_T^2}+\sqrt{\frac{1}{J-1}\sum_{j=1}^{J-1}||d_{f^*}(j,\cdot)-P(j,\cdot,f_0)||_T^2}\right)
\end{aligned}$$

where the third inequality holds similarly with (45), the fourth inequality holds by the triangle inequality, and the last inequality is due to the Cauchy-Schwarz inequality. Therefore, from Assumption 5.3 and the first result of Theorem 5.2, it follows that

$$\sum_{r=1}^{R}\left|\widehat{f}(r)-f^*(r)\right|=O\left(\sqrt{R}\sqrt{\frac{R\cdot\log(RT)}{T}}\right)=O\left(R\sqrt{\frac{\log(RT)}{T}}\right).$$

References

[1] Ackerberg, Daniel A. 2001. "A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation." working paper.

[2] Amemiya, T. (1983), "Non-Linear Regression Models," *Handbook of Econometrics* I, edited by Z. Griliches and M.D. Intrilligator, 333-389.

[3] Andrews, D.K.W. (1994), "Empirical Process Methods in Econometrics," *Handbook of Econometrics* IV, edited by R.F. Engle and D.L. McFadden, 2247-2294.

[4] Andrews D.K.W. (2002a), "Estimation When a Parameter is on a Boundary," *Econometrica* 67, 1341-1383.

[5] Andrews D.K.W. (2002b), "Generalized Method of Moments Estimation When a Parameter is on a Boundary," *Journal of Business & Economics Statistics* 20-4, 530–544.

[6] Andrews D.K.W and P. Guggenberger (2005), "Hybrid and Size-corrected Subsample Methods," Cowles Foundation Discussion Paper No. 1606, Yale University.

[7] Bajari, Patrick and C. Lanier Benkard (2005). "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach." The Journal of Political Economy.

[8] Bajari, P., J.T. Fox, and S. Ryan (2007), "Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients," *AER Papers and Proceedings.* May, 459–463.

[9] Baraud, Y. (2002), "Model Selection for Regression on a Random Design," *ESAIM Probability & Statistics 7*, 127-146

[10] Barbe, P. (1998) "Statistical analysis of mixtures and the empirical probability measures," Acta Applicandae Mathematicae, 50(3), 253–340.

[11] Berry, S., J. Levinsohn, and A. Pakes, "Automobile Price in Market Equilibrium", Econometrica (63), July 1995.

[12] Berry, Steve, Oliver B. Linton and Ariel Pakes (2004) "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," Review of Economic Studies 71 (3), 613–654.

[13] Blundell, R., X. Chen, and D. Kristensen (2007), "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," forthcoming in *Econometrica*.

[14] Böhning, D. "Convergence of Simar's Algorithm for Finding the Maximum Likelihood Estimate of a Compound Poisson Process", The Annals of Statistics, 10(3), 1006–1008. 1982.

[15] Boyd, J. Hayden and Robert E. Mellman. 1980. "The Effect of Fuel Economy Standards on the U.S. Automobile Market: An Hedonic Demand Analysis." *Transportation Research Part A: General*, 14A: 367–378.

[16] Briesch, R.A., P.K. Chintagunta, and R.L. Matzkin (2007), "Nonparametric Discrete Choice Models with Unobserved Heterogeneity," Working paper.

[17] Bunea, F., A.B. Tsybakov, and M.H. Wegkamp (2006), "Aggregation and Sparsity via l_1 Penalized Least Squares," Proceedings of the Annual Conference on Learning Theory, Lecture Notes in Artificial Intelligence, Springer, 379-391.

[18] Cardell, N. Scott and Frederick C. Dunbar. 1980. "Measuring the societal impacts of automobile downsizing." *Transportation Research Part A: General*, 14A: 423–434.

[19] Chintagunta, Pradeep K., Dipak C. Jain and Naufel J. Vilcassim. "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data." *Journal of Marketing Research*, 28(4): 417–428.

[20] Cramér, H. and H. Wold. 1936. "Some Theorems on Distribution Functions," Journal of the London Mathematical Society, s1-11(4), 290–294.

[21] D'Haultfoeuille, Xavier. 2007. "On the Completeness Condition in Nonparametric Instrumental Problems", working paper.

[22] Erdem, Tulin, Susumu Imai and Michael P. Keane. 2003. "Brand and Quantity Choice Dynamics Under Price Uncertainty." Quantitative Marketing and Economics, 1(1), 5–64.

[23] Follmann, Dean A. and Diane Lambert. 1989. "Generalized Logistic Regression by Nonparametric Mixing." *Journal of the American Statistical Association*, 81(393): 295–300.

[24] Fox, Jeremy T. and Amit Gandhi. 2007. "Identifying Heterogeneity in Discrete Choice, Selection and Other Economic Models." working paper.

[25] Fox, Jeremy T. and Che-Lin Su. 2007 "Improving the Numerical Performance of BLP Structural Demand Estimators," working paper.

[26] Geweke, J. "Exact Inference in the Inequality Constrained Normal Linear Regression Model," " Journal of Applied Econometrics, Vol. 1, No. 2. (Apr., 1986), pp. 127-141.

[27] Heckman, J. and Singer, B. " Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* 52(2), 271-320 .

[28] Hendel, Igal and Aviv Nevo. "Measuring the Impact of Sales and Consumer Inventory Behavior." Econometrica, 2006.

[29] W. Hoeffding (1963), "Probability Inequalities for Sums of Independent Random Variables," *Journal of the American Statistical Association*, 58, 13-30.

[30] Huang, J., J. Horowitz, and S. Ma (2007), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," Working paper.

[31] Ichimura, Hidehiko and T. Scott Thompson (1998), "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution,", Journal of Econometrics, 86(2), 269–295.

[32] Judge, G.G. and Takayama, T. "Inequality Restrictions in Regression Analysis," Journal of the American Statistical Association, Vol. 61, No. 313. (Mar., 1966), pp. 166-181.

[33] Laird, Nan (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution", Journal of the American Statistical Association, Vol. 73, No. 364, pp. 805–811.

[34] Leeb, H. and B.M. Pötscher (2005), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," Cowles Foundation Discussion Paper No. 1500, Yale University.

[35] Lehmann, E.L. (1986), *Testing Statistical Hypothesis*, 2nd Ed. Wiley.

[36] Lehmamm, E.L. and J.P. Romano (2005), *Testing Statistical Hypotheses*, 3rd Ed. Springer.

[37] Li, Jonathan Q. and Andrew R. Barron (2000), "Mixture density estimation", Advances in Neural Information Processing Systems, Vol. 12, pp. 279–285.

[38] Liew, C.K. "Inequality Constrained Least-Squares Estimation," Journal of the American Statistical Association, Vol. 71, No. 355. (Sep., 1976), pp. 746-751.

[39] Lindsay, B.G. (1983) "The Geometry of Mixture Likelihoods: A General Theory", The Annals of Statistics, 11(1), 86–94.

[40] Manski, Charles F. (1975) "Maximum Score Estimation of the Stochastic Model of Choice", Journal of Econometrics, 3(3), 205–228.

[41] C.F. Manski (2007), "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48-4, 2007.

[42] McFadden, Daniel and Kenneth Train. 2000. "Mixed MNL models for discrete response." *Journal of Applied Econometrics*, 15(5): 447–470.

[43] Nevo, A "A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand," Journal of Economics & Management Strategy, 9(4), 513-548, 2000.

[44] Nevo, Aviv. 2001, "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica*, 69(2): 307–342.

[45] Newey, W.K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

[46] Petrin, Amil. "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110:705-729, 2002.

[47] Petrin, Amil and Kenneth Train, "Control Function Corrections for Omitted Attributes in Differentiated Products Markets," 2006 working paper

[48] Rossi, Peter E., Greg M. Allenby, and Robert McCulloch. 2005. *Bayesian Statistics and Marketing*. West Sussex: John Willy & Sons.

[49] Roueff, Francois and Tobias Rydén (2005). "Nonparametric estimation of mixing densities for discrete distributions," Annals of Statistics, 33(5), 2066–2108.

[50] Rust, John. 1987. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica*, 55(5): 999–1033.

[51] Rust, John. "Structural Estimation of Markov Decision Processes." In Handbook of Econometrics, vol. 4, edited by Robert F. Engle and Daniel L. McFadden. Amsterdam: North-Holland, 1994.

[52] Shohat, J.A. and Tamarkin, J.D. (1943), *The Problem of Moments*, American Mathematics Society, Providence, RI.

[53] Teicher, H. (1963), "Identifiability of Finite Mixtures," *Annals of Mathematical Statistics*, 34, 1265-1269.

[54] Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B.* 58, 267-288.

[55] Train, Kenneth. 2003. *Discrete Choice Methods with Simulation.* Cambridge: Cambridge University Press.

[56] Verbeek, J.J., N. Vlassis and B. Kröse (2003), "Efficient Greedy Learning of Gaussian Mixture Models," Neural Computation, Vol. 15, pp 469–485.

[57] Wolak, F. "Testing inequality constraints in linear econometric models," Journal of Econometrics, Elsevier, vol. 41(2), pages 205-235, June 1989.

[58] V.V. Yurinskii (1976), "Exponential Inequalities for Sums of Random Vectors," *Journal of Multivariate Analysis*, 6, 473-499.

[59] Zeithammer, Robert and Peter Lenk (2006), "Bayesian Analysis of Multivariate Normal Models when Dimensions are Absent", Quantitative Marketing and Economics, Vol 4, No 3, pp 241–265.