The random coefficients logit model is identified

Jeremy T. Fox [a,*], Kyoo il Kim [b], Stephen P. Ryan [c], Patrick Bajari [d]

[a] *University of Michigan & NBER, United States*
[b] *University of Minnesota, United States*
[c] *MIT & NBER, United States*
[d] *University of Minnesota & NBER, United States*

ARTICLE INFO

ABSTRACT

The random coefficients multinomial choice logit model, also known as the mixed logit, has been widely used in empirical choice analysis for the last thirty years. We prove that the distribution of random coefficients in the multinomial logit model is nonparametrically identified. Our approach requires variation in product characteristics only locally and does not rely on the special regressors with large supports used in related papers. One of our two identification arguments is constructive. Both approaches may be applied to other choice models with random coefficients.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

One of the most commonly used models in applied choice analysis is the random coefficients logit model, also known as the mixed logit, which describes choice between one of a finite number of competing alternatives. Domenich and McFadden (1975), Heckman and Willis (1977) and Hausman and Wise (1978) introduced flexible specifications for discrete choice models, while the random coefficients logit model was proposed by Boyd and Mellman (1980) and Cardell and Dunbar (1980). Currently, the random coefficients logit model is widely used to model consumer choice in environmental economics, industrial organization, marketing, public economics, transportation economics and other fields.

In the random coefficients logit, each consumer can choose between $j = 1, \ldots, J$ mutually exclusive inside goods and one outside good (good 0). The exogenous variables for choice j are in the $K \times 1$ vector $x_j = (x_{j,1}, \ldots, x_{j,K})'$. In the example of demand estimation, x_j might include the product characteristics, the price of good j and the interactions of product characteristics with the demographics of agent i. A source of variation in product characteristics is the differences in characteristics across markets, which could be indexed by t. So potentially x_j can depend on consumers i and markets t, but we suppress these subscripts for transparency. In this paper, for some results we assume that the support of x_j includes the 0 vector, which can occur by centering each element of x_j around constants common across the inside goods. For example, for each k we may redefine $\tilde{x}_{j,k} =$

$x_{j,k} - E[x_{j,k}]$, where the mean is taken over the inside goods and across markets or consumers when the exogenous variables vary across markets or consumers. This location normalization for the covariates is not without loss of generality in a model, like the logit, where the additive error is treated parametrically. However, one normalization is just as arbitrary as any other. Let $x = (x'_1, \ldots, x'_J)$ denote the stacked vector of all the x_j.

Each consumer i has a preference parameter β_i, which is a vector of K marginal utilities that gives i's preferences over the K product characteristics. There is also a homogeneous term for each choice j, denoted as α_j. Each α_j could be an intercept common to product j or a term that captures product characteristics without random coefficients (homogeneous coefficients), as in $\alpha_j = \alpha + w'_j \gamma_w$. Here α is a common intercept for the inside goods contributing to the utility of purchasing an inside good instead of the outside good, γ_w is a vector of parameters on the characteristics in the vector w_j, and $w = (w'_1, \ldots, w'_J)$ denotes the stacked vector of other product characteristics. Let α^J be the vector of the J α_j's. Agent i's utility for choice j is equal to

$$u_{i,j} = \alpha_j + x'_j \beta_i + \epsilon_{i,j}. \tag{1}$$

The outside good has a utility of $u_{i,0} = 0 + \epsilon_{i,0}$, as a location normalization for the utilities. The type I extreme value distribution gives the scale normalization for utility values.

The logit model is defined when the errors $\epsilon_{i,j}$ are i.i.d. across choices and each error has the type I extreme value distribution, which has a cumulative distribution function of $\exp\left(-\exp\left(-\epsilon_{i,j}\right)\right)$. The random coefficients logit arises when β_i varies across the population, with unknown distribution $F(\beta_i)$. The unknown objects of estimation are the distribution $F(\beta_i)$ and the homogeneous coefficients (α, γ'_w) or α^J. Under the standard

---

* Correspondence to: University of Michigan, 48109-1220 Ann Arbor, MI, United States. Tel.: +1 734 330 2854; fax: +1 734 274 2331.
E-mail address: jtfox@umich.edu (J.T. Fox).

assumption that β_i is independent of x_j (we discuss endogeneity below), utility maximization leads to a choice probability for the good j,

$$\Pr\left(j \mid x; F, \alpha^J\right) \equiv \int \frac{\exp\left(\alpha_j + x'_j \beta_i\right)}{1 + \sum_{j'=1}^{J} \exp\left(\alpha_{j'} + x'_{j'} \beta_i\right)} \, dF\left(\beta_i\right). \tag{2}$$

This specification is popular with empirical researchers because the resulting choice probabilities are relatively flexible. In terms of modeling own- and cross-price elasticities, the random coefficients logit model allows products with similar x's to be closer substitutes, which the logit model without random coefficients does not allow.

In this paper, we prove that the distribution $F(\beta)$ is nonparametrically identified, in the sense that the true $\left(F^0, \alpha^{J,0}\right)$ is the only pair $\left(F, \alpha^J\right)$ that solves $\Pr\left(j \mid x, w\right) \equiv \Pr\left(j \mid x; F, \alpha^J\right)$ in (2) for all j and all (x, w), where $\Pr\left(j \mid x, w\right)$ denotes the population choice probabilities. We first recover the homogeneous terms, α^J (or the fixed parameters α and γ_w) in (2). We then provide two identification arguments for the distribution of random coefficients, one of which is constructive and the other of which is non-constructive. The non-constructive identification theorem leverages results from Hornik (1991) on function approximation and Stinchcombe and White (1998) on consistent specification testing. The theorem requires variation in x in only an open set.

Our other identification argument is constructive. We demonstrate how to iteratively find all moments of β, which is sufficient to identify the distribution F^0 within the class of distributions that are uniquely determined by all of their moments. This class is the set of probability distribution functions that satisfy the Carleman condition, which we review below.

Both the constructive and non-constructive proof strategies are not unique to the logit: they could be applied to identify the distribution of heterogeneity in many differentiable economic models where covariates enter as linear indices. We outline the main theorems using generic notation and verify their conditions for the multinomial logit model.

While one of our identification approaches is constructive, we do not recommend that empirical researchers adopt an analog estimator to our identification argument. Instead, we suggest empirical users adopt one of several nonparametric mixtures estimators available in the literature. Before this paper, no one had formally proved that these mixtures estimators consistently estimated the true F^0 in the random coefficients logit. This lack of a complete consistency proof arises because showing that the density F^0 is nonparametrically identified is a necessary component for any consistency proof for a nonparametric estimator of F^0. We introduce a computationally simple, nonparametric sieve estimator for F^0 in Fox et al. (forthcoming) and Fox and Kim (2011) for general mixtures models. We prove that our estimator is consistent in the Lévy–Prokhorov metric on distributions under the maintained assumption that the model is identified. This identification theorem therefore completes our proof of consistency for the estimator of the random coefficients logit, based on Corollary 1 in Fox and Kim. Alternative nonparametric estimators used for the random coefficients logit include the Bayesian MCMC estimators in Rossi et al. (2005) as well as Burda et al. (2008) and the EM algorithm used in Train (2008). These works do not discuss consistency or identification. However, the identification theorem here would be a building block to proving the consistency of the estimators of $\left(F, \alpha^J\right)$ in these other procedures.

The proof of identification is comforting to empirical researchers. Prior to our theorem, it was not known whether variation in x was sufficient to identify the pair $\left(F^0, \alpha^{J,0}\right)$. One possibility was that the normality assumptions typically imposed on F^0 were

crucial to identification. We show that indeed the random coefficients logit model is identified, which provides a more solid econometric foundation for its application in applied microeconomics.

The paper is organized as follows: Section 2 discusses the related literature, Section 3 introduces notation for a generic mixtures model, Section 4 states the general non-constructive identification result, Section 5 states the general constructive identification result, Section 6 shows how the results apply to the random coefficients logit, Section 7 explores extensions, and Section 8 concludes.

2. Related literature

There is a growing literature on the identification of binary and multinomial choice models with unobserved heterogeneity. These papers differ in the set of assumptions made in order to obtain identification results and also differ in the objects of interest in identification.

Ichimura and Thompson (1998) study the case of binary choice: one inside good ($J = 1$) and one outside good. This restriction makes their method inapplicable for most empirical applications of demand analysis, which study markets with two or more inside goods. Ichimura and Thompson identify the cumulative distribution function (CDF) of, in our notation, $\left(\beta, \epsilon_{i,1} - \epsilon_{i,0}\right)$. They use a theorem due to Cramér and Wold (1936) and do not exploit the structure of the extreme value assumptions on the $\epsilon_{i,1}$ and $\epsilon_{i,0}$. Consequently, they need stronger assumptions: (1) a monotonicity assumption (sign restriction) on one of the K components of β ($\beta_{i,k} > 0 \,\forall i$) and (2) a full support assumption for all K elements of x_1. Similar assumptions will appear in many of the papers below. Gautier and Kitamura (2009) provide a computationally simple estimator and some alternative identification arguments for the same binary choice model as Ichimura and Thompson. To our knowledge, no one has generalized Ichimura and Thompson to the case of multinomial choice.

We will refer to a regressor whose (random) coefficient has a sign restriction and that has full support as a "special regressor". The sign restriction for a special regressor may not be too restrictive because it is testable and the sign can be determined from a pre-model analysis as described in Lewbel (2000) and his other related work. Magnac and Maurin (2007) and Kahn and Tamer (2010), however, argue that identification using the large support condition and a conditional mean restriction, as in Lewbel (2000), critically relies on the tail behavior of the distribution of the special regressor, as is also the case in "identification at infinity" in selection models (Andrews and Schafgans, 1998).

Lewbel (2000) provides an identification argument that relies on a large-support special regressor, but which allows for discrete elements of x_j, relaxes the independence between x and ϵ, and does not rely on distributional assumptions of the error term. Lewbel's approach identifies the means of the random coefficients and the conditional distribution of the composite error $x'_j \left(\beta_i - E\left[\beta_i \mid x\right]\right) + \epsilon_{i,j}$. However, identifying the distribution of the composite error does not identify the joint distribution of the random coefficients, which is the key object of interest in our paper.

Briesch et al. (2010) study the identification of a discrete choice model where the payoff to choice j is $V\left(j, z_j, s_i, \omega_i\right) - r_j + \epsilon_{i,j}$, where V is an unknown, nonparametric function common to all consumers, z_j are observed product characteristics, s_i are observed consumer characteristics, r_j is a large support special regressor with a sign restriction as in Ichimura and Thompson, $\epsilon_{i,j}$ is an additive error and ω_i is a scalar unobservable that enters the utility functions for all J choices. There are a variety of other restrictions. Their model does not nest the random coefficients logit. Matzkin (2007, page 101) extends these results to the utility function for choice j of $r_j + \sum_{k=1}^{K} m_k\left(z_{j,k}, \omega_{i,j,k}\right) + \epsilon_{i,j}$, where

each function $m_k (\cdot, \cdot)$ is treated nonparametrically, each choice j has its own random coefficients $\omega_{i,j} = (\omega_{i,j,1}, \ldots, \omega_{i,j,K})$, the random coefficients are independent across choices, and r_j is a special regressor. The special regressor and the independence of random coefficients across choices mean that the standard random coefficients logit model is not nested in the formulation of Matzkin.

Subsequent to the circulation of our constructive identification theorem, Berry and Haile (2010) and Fox and Gandhi (2010) introduced identification arguments for multinomial choice models without the type I extreme value distribution or additive errors. Both Berry and Haile and Fox and Gandhi need a monotonicity assumption on one of the K components of β ($\beta_{i,k} > 0 \ \forall i$) and (for point instead of set identification) a full support assumption on the corresponding k-th component $x_{j,k}$, for all choices $j \in J$. Berry and Haile identify the conditional-on-x distribution of utility values $G\left(u_{i,0}, u_{i,1}, \ldots, u_{i,J} \mid x\right)$ and not $F(\beta)$. Knowledge of the full structural model, in the logit case $F(\beta)$, is necessary for welfare analysis, for example to construct the distribution of welfare gains between choice situations x^1 and x^2, or some aggregation of welfare gains over individuals $H\left(\Delta_i u \mid x^1, x^2\right)$, where

$$\Delta_i u = \max_{j \in J \cup \{0\}} u_{i,j}\left(x^1\right) - \max_{j \in J \cup \{0\}} u_{i,j}\left(x^2\right),$$

where $u_{i,j}\left(x^l\right)$ is just the realized utility value (1) for $x^l = \left(x_1^{l\prime}, \ldots, x_J^{l\prime}\right)$. Fox and Gandhi do identify the full structural model, in that they identify a distribution D over J utility functions (not utility values) of x, as in $D\left(u_{i,1}(x), \ldots, u_{i,J}(x)\right)$, where $u_{i,j}(x)$ is a complete function that describes utility values for choice j at all x. Again, like all results other than ours, Fox and Gandhi rely on monotonicity and large support assumptions for a special regressor.

Compared to this other literature, our main distinguishing feature is that we exploit the logit distributional assumptions on the $\epsilon_{i,j}$. This corresponds to empirical practice: the random coefficients logit is a popular specification in applied work. McFadden and Train (2000, Theorem 1) present an approximation theorem using the random coefficients logit as the approximating class, although the theorem requires great flexibility in the choice of the product characteristics x_j in the random coefficients estimation as a function of some smaller set of underlying true product characteristics. McFadden and Train do not study identification.

Our results contribute to the literature on the identification of discrete choice models by demonstrating that large support and monotonicity restrictions are not required for identification if the logit error structure is used, as is common in empirical work. In our opinion, the main concern with the special regressor assumption is the requirement for large support. Large supports are sometimes but not often found in typical datasets used in demand estimation. Price may be a special regressor; certainly the assumption of monotonicity on price is rarely controversial. In an output market, prices are some markup over cost, and cost rarely moves more than a factor of, say, five (think oil price fluctuations). As we show, the parametric assumption on the distribution of the choice-specific errors does away with the need for large support assumptions. The entire distribution of random coefficients can be identified using only local variation in characteristics. In subsequent and at present in-progress work, Chiappori and Komunjer (2009) present preliminary results for achieving a weakening of conditions on special regressors without parametric assumptions in the multinomial choice model.

Our two identification approaches may be applied to choice models with random coefficients other than the logit, as we describe below. Our paper, like many of the papers in the literature, focuses on continuous covariates in x. Lewbel (2000) allows for discrete covariates other than in the special regressor

and shows the identification of, in our notation, the distribution of $x_j'(\beta - E[\beta|x]) + \epsilon_{i,j}$ conditional on x. He does not explore identification of an unconditional joint distribution $F(\beta)$. Knowledge of the unconditional distribution $F(\beta)$ is necessary for some uses of structural models, including prediction of demand and x's not in the support of the original data (the new goods problem).

All of our arguments can be made conditional on the values of discrete covariates, but we do not explore identifying a distribution of random coefficients on discrete covariates. No other paper has identified such a distribution, either. Therefore, our results do not allow the random coefficients logit to be a flexible error components model for parameterizing the correlation between products grouped into various nests. Unlike the subset of the literature that is nonparametric on the contribution of x to the utility of each choice, we follow Ichimura and Thompson (1998) and the widespread empirical literature using the random coefficients logit and focus on the linear index $x_j'\beta_i$.

3. Generic model

To deliver the key idea behind our identification strategy, we shall first consider an abstract model that includes the random coefficients logit with fixed intercepts α as a special case. The econometrician observes $K < \infty$ covariates $x = (x_1, \ldots, x_K)'$ and the probability of some discrete outcome, $P(x)$. Here $P(x)$ denotes the conditional choice probability of a particular outcome. For a model with a more complex outcome (including a continuous outcome y), we can always consider whether some event, say $y < \frac{1}{2}$, happened or did not happen. For a multinomial choice, the event could be picking choice j or picking any other choice. $P(x)$ is the probability of the event happening. The regressor vector x is independent of β.

Let $g\left(\alpha, x'\beta\right)$ be the probability of an agent with characteristics β taking the action. In our framework, the researcher specifies $g\left(\alpha, x'\beta\right)$. A special case is $g\left(\alpha, x'\beta\right) = g\left(\alpha + x'\beta\right)$. Our goal is to identify the distribution function $F(\beta)$ in the equation

$$P(x) = P\left(x, F, \alpha\right) \equiv \int g\left(\alpha, x'\beta\right) dF(\beta). \tag{3}$$

Identification means that a unique (F, α) solves this equation for all x. This is the definition of identification used in the statistics literature (Teicher, 1963).

We prove the main theoretical results for the case where the true values α^0 of the homogeneous parameters α are identified in a first stage using an auxiliary argument, so we can write $g\left(x'\beta\right) = g\left(\alpha^0, x'\beta\right)$, with an abuse of notation. Typically, this auxiliary argument will involve the point $x = 0$, as then (3) becomes $P(0) = P(0, F, \alpha) = g(\alpha, 0)$ and in some models we can set $\alpha^0 = g^{-1}(P(0))$. The point $x = 0$ will otherwise not be needed to be in the support of x for the general result. We rewrite the model as

$$P(x, F) \equiv \int g\left(x'\beta\right) dF(\beta). \tag{4}$$

Let $\mathcal{B} \subseteq \mathbb{R}^K$ be the support of the random coefficients and let $\mathcal{F}(\mathcal{B})$ be the set of all distributions on that support. Let $\mathcal{X} \subseteq \mathbb{R}^K$ be the support of the covariates. Let F^0 be the true distribution. Then we have $P(x) = P\left(x, F^0\right)$ in the support of x, $\mathcal{X}$.

4. Non-constructive identification

We first present a non-constructive identification theorem and, in a later section, turn to constructive identification. Our constructive identification section is written to be self-contained, so a reader can skip this section if not interested.

Definition 1. The distribution $F^0 \in \mathcal{F}(\mathcal{B})$ is *uniformly identified over choices of* $(\mathcal{B}, \mathcal{X}_0)$ if for any $F^1 \in \mathcal{F}(\mathcal{B})$, $F^1 \neq F^0$, there exists

$\mathcal{X}_0^1 \subset \mathcal{X}_0$ such that $P\left(x, F^0\right) - P\left(x, F^1\right) \neq 0$ for all $x \in \mathcal{X}_0^1$ for any choices of the support of random coefficients $\mathcal{B}$ and the subset of the support of covariates $\mathcal{X}_0 \subset \mathcal{X}$, where $\mathcal{B}$ is compact and $\mathcal{X}_0$ is a nonempty open set.

Note that in this definition the set $\mathcal{X}_0^1$ can vary depending on the alternative distribution F^1. When the measure of the set $\mathcal{X}_0^1$ is strictly positive, identification holds with positive probability. We do not assume the researcher knows $\mathcal{B}$, other than that it is compact. The "nonempty interior" assumption below does rule out covariates with discrete support, as we discussed earlier. All arguments in this paper can be made conditional on the values of discrete characteristics. Otherwise and as a referee suggests, one may take a set identification approach when x takes values on a fine grid. However, this is beyond the scope of this paper. For explicitness, we emphasize the assumption on the covariates implied in the previous definition.

Assumption 2. 1. The support of the independent variables, $\mathcal{X} \subseteq \mathbb{R}^K$ includes a nonempty open set.
2. Let all elements of x be continuous.

Assumption 2.1 will be violated if x includes higher order terms of x_k's or interactions of x_k's (e.g., $x_2 = x_1^2$ or $x_3 = x_2 \cdot x_1$) in its elements. Therefore our identification results do not allow for those terms in the model.

Definition 3. The function $g(z)$ is *real analytic* at $c \in \mathcal{Z} \subseteq \mathbb{R}$ whenever it can be represented as a convergent power series, $g(z) = \sum_{d=0}^{\infty} a_d (z - c)^d$, for a domain of convergence around c. The function $g(z)$ is *real analytic* on an open set $\mathcal{Z} \subseteq \mathbb{R}$ if it is real analytic at all arguments $z \in \mathcal{Z}$.

Similarly (Definition 2.2.1 in Krantz and Parks (2002)) a function $\Delta(x)$, with domain an open subset $\mathcal{T} \subseteq \mathbb{R}^K$ and range $\mathbb{R}$, is called (multivariate) real analytic on $\mathcal{T}$ if for each $x \in \mathcal{T}$ the function $\Delta(\cdot)$ may be represented by a convergent power series in some neighborhood of x.

Note that the domain of the real analytic function $g(\cdot)$ is a subset of $\mathbb{R}$. A real analytic function on a real domain may not be analytic on a complex domain. Our non-constructive identification theorem follows.

Theorem 4. *Assume that the vector α^0 is identified using some auxiliary argument and let Assumption 2 hold. Let $\mathcal{B}$ be compact and $\mathcal{X}_0$ be a nonempty open subset of $\mathcal{X}$. The distribution $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over choices of $(\mathcal{B}, \mathcal{X}_0)$ if the function $g(\cdot)$ is real analytic, bounded, nonconstant and satisfies $g(0) \neq 0$.*

4.1. Lemmas that factor into the proof of Theorem 4

The proof of Theorem 4 is a consequence of two lemmas. We state the lemmas here for those interested in the logic behind the identification theorem.

We say that $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over choices of $\mathcal{B}$, holding a set of values of x, $\mathcal{T}_0$, fixed, whenever $P\left(x, F^0\right) - P\left(x, F^1\right) \neq 0$ for all $x \in \mathcal{T}_0^1 \subset \mathcal{T}_0$ for any $F^1 \in \mathcal{F}(\mathcal{B})$, $F^1 \neq F^0$ and for any compact choice for the support of the random coefficients, $\mathcal{B}$. In this definition, $\mathcal{T}_0$ does not have to be included in the support of x. $P\left(x, F^0\right)$ and $P\left(x, F^1\right)$ for $x \in \mathcal{T}_0 \not\subseteq \mathcal{X}$ are well defined from the model in (4) and then we have $P(x) = P\left(x, F^0\right)$ only for $x \in \mathcal{X}$ if F^0 is the true distribution.

Lemma 5. *Assume that the vector α^0 is identified using some auxiliary argument. Let $g(\cdot)$ be real analytic and let a set of x, $\mathcal{T}$, contain a nonempty open set. The distribution $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over choices of $(\mathcal{B}, \mathcal{T}_0)$, with $\mathcal{B}$ compact, with nonempty open sets $\mathcal{T}_0 \subset \mathcal{T}$ if and only if $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over compact choices of $\mathcal{B}$, for at least one fixed $\mathcal{T}_0 \subseteq \mathcal{T}$.*

The lemma and its proof (in our Appendix for completeness) are inspired by Theorem 3.8 in Stinchcombe and White (1998), a paper on consistent specification testing. The content of Lemma 5 is that identification for any choice of nonempty open set $\mathcal{T}_0$ automatically holds if identification is checked for one, likely convenient choice of $\mathcal{T}_0$. The most convenient choice of $\mathcal{T}_0$ is the one with the widest variation, or $\mathcal{T}_0 = \mathcal{T} = \mathbb{R}^K$ as in Lemma 6 below.

The key idea of Lemma 5 is as follows. When $g(\cdot)$ is real analytic and $\mathcal{B}$ is compact, the integral of the form $\Delta(x) \equiv \int g\left(x'\beta\right) d(F^1(\beta) - F^0(\beta))$ itself is real analytic. If a real analytic function is equal to 0 on an open set, it equals 0 everywhere.[1] Therefore if $\Delta(x)$ is 0 on an open set T_0, it must be zero everywhere on $\mathcal{T}$. This implies, as we formalize in the proof, that if identification holds on an open set T_0, it should also hold everywhere.

Lemma 6. *Assume that the vector α^0 is identified using some auxiliary argument. Let $g(\cdot)$ be bounded and nonconstant and satisfy $g(0) \neq 0$. Then the distribution $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over choices of $\mathcal{B}$ for the choice $\mathcal{T} = \mathbb{R}^K$.*

Lemmas 5 and 6 together imply Theorem 4. Lemma 6 shows that the model (4) identifies F^0 against any $F^1 \neq F^0$ with $\mathcal{T}_0 = \mathcal{T} = \mathbb{R}^K$ and Lemma 5 says then identification should hold with any nonempty open set $\mathcal{T}_0 \subseteq \mathcal{T}$. Therefore, in Theorem 4 we conclude that the F^0 that solves $P(x) = P(x, F)$ in (4) is identified with any nonempty open set $\mathcal{X}_0 \subset \mathcal{X}$, noting that we have $P(x) = P(x, F^0)$ in the support of x. Lemma 5 shows the role of $g(\cdot)$ being real analytic for identification: it removes the full support condition.

Lemma 6 follows from Theorem 5 in Hornik (1991) with very minor modifications (see its proof for completeness in our Appendix). Hornik is a paper on functional approximation using a function that takes a linear index. To understand the connection, note that if F^0 is not identified on $\mathcal{T} = \mathbb{R}^K$ we must have $F^1 \neq F^0$ such that

$$0 = P(x, F^1) - P(x, F^0)$$

$$= \int_{\mathcal{B}} g(x'\beta) d(F^1(\beta) - F^0(\beta)) \quad \text{for all } x \in \mathcal{T}. \tag{5}$$

In the context of our identification problem, Hornik's Theorem 5 says that as long as $g(\cdot)$ is bounded and nonconstant, only $F^1 = F^0$ satisfies (5) (i.e. (5) implies $F^1 = F^0$) and hence identification holds. Hornik called a function $g(\cdot)$ that has this *identification* property *discriminatory*.

Other than the requirement that $g(0) \neq 0$, Lemma 6 would be exactly the same as the statement of Theorem 5 of Hornik (in our identification context) if our model was $g\left(x_1'\beta + x_2\right)$, where in a study of identification x_2 is a special regressor with large support and known sign. However, in our identification problem it suffices to modify Hornik's general result (and its proof) so that x_2 is only used to establish that $g(x_2) \neq 0$ for some $x_2 \in \mathbb{R}$ (because g is nonconstant), which we replace with the requirement that $g(0) \neq 0$. We do not need data on $x = 0$ to establish that $g(0) \neq 0$; indeed $P(0, F) = g(0)$ is known from the choice of g (and the identification of α), as it is a trivial function of F. To our knowledge, no other paper has linked the result on function approximation in Hornik to the identification of distributions of heterogeneity.

---

[1] Suppose $h(z)$ is a real analytic function on an open interval $\mathcal{Z}$ and equal to zero on an open sub-interval $\mathcal{Z}_0 \subseteq \mathcal{Z}$ containing z_0. Then we have $h^{(l)}(z_0) = 0$ for any l-th order derivative of h. Now set $\tilde{\mathcal{Z}} = \mathcal{Z} \cap \{z: h^{(l)}(z) = 0 \text{ for } l = 0, 1, 2, \ldots\}$. By continuity, $\tilde{\mathcal{Z}}$ is closed in the relative topology of $\mathcal{Z}$, while by assumption on $\mathcal{Z}$, $\tilde{\mathcal{Z}}$ is open. Thus, by the connectedness of $\mathcal{Z}$, we have $\tilde{\mathcal{Z}} = \mathcal{Z}$. Therefore if an analytic function $h(z)$ is zero on an open interval, it is equal to zero everywhere (Corollary 1.2.6 in Krantz and Parks (2002)). Consider a simple example $h(z) = \frac{\exp(az) + b \exp(cz)}{1 + \exp(z)}$ and note that $h(z)$ is real analytic because sums or reciprocal of real analytic functions are real analytic. Assume $h(z_1) = h(z_2) = 0$ for $z_1 \neq z_2 \in \mathcal{Z}$. Then simple algebra yields $a = c$ and $b = -1$, implying $h(z) = 0$ everywhere.

5. An alternative, constructive approach

The previous identification approach is fairly general but is not constructive. In other words, the identification argument implies that $P\left(x, F^0\right) - P\left(x, F^1\right) \neq 0$ for a set of x with a nonempty interior (positive probability) but does not give a procedure to construct F^0. Here we give such a constructive identification argument. This section is self-contained in that the results do not refer to the non-constructive identification arguments. The downside of the constructive arguments is that an open set around the point $x = 0$ will play a special role in the identification of the distribution of random coefficients, while the point $x = 0$ only possibly played the role of identifying homogeneous parameters α in the prior, non-constructive argument. On the other hand, compactness of the parameter space $\mathcal{B}$ will not be imposed, as it was in the non-constructive argument. Indeed, $\mathcal{B}$ can equal $\mathbb{R}^K$. The function $g(\cdot)$ will not need to be real analytic. As before, the homogeneous parameters α can be identified in a first stage.

Assumption 7. The absolute moments of $F(\beta)$, given by $m_l = \int \|\beta\|^l \, dF(\beta)$, are finite for $l \geq 1$ and satisfy the Carleman condition: $\Sigma_{l \geq 1} m_l^{-1/l} = \infty$.

A distribution F satisfying the Carleman condition is uniquely determined by its moments (Shohat and Tamarkin, 1943, p. 19). The Carleman condition is weaker than requiring the moment generating function to exist. The Carleman condition gives uniqueness for distributions with unrestricted support. If the support of F is known and compact, uniqueness follows without the Carleman condition. Let $g^{(l)}(c)$ be the l-th derivative of $g(c)$ evaluated at c.

Assumption 8. 1. $g(c)$ is infinitely differentiable on an open set $\mathcal{C} \subset \mathbb{R}$ that includes $c = 0$.
2. $g^{(l)}(0)$ is nonzero and finite for all $l \geq 1$.

For a (non-probability) example, if $g(c) = D \cdot \exp(c)$, then Assumption 8 is satisfied because $g^{(l)}(0) = D$ for all l. The logit choice probabilities will be shown to be infinitely differentiable. We will investigate to what extent Assumption 8.2 holds for logit choice probabilities when we turn from the generic model to the logit model. If $g(\cdot)$ is a polynomial function of any finite degree, g does not satisfy Assumption 8.2 because its derivative becomes zero at a certain point. Also, a particular polynomial of order p could potentially have all of its derivatives of lower and higher order than order p be 0 at $c = 0$. For polynomials where only derivatives of order higher than v are 0 at $c = 0$, we identify the distribution of β up to the v-th moment.

We will maintain the support assumption for the regressors that was earlier used for the non-constructive results.

Assumption 9. 1. The support of the independent variables, $\mathcal{X} \subseteq \mathbb{R}^K$ includes a nonempty open set.
2. Let all elements of x be continuous.

The key limitation of the constructive identification theorem is its reliance on regressor variation in an open set around the point $x = 0$.

Assumption 10. The support $\mathcal{X}$ contains an open set surrounding $x = 0$.

Recall from the introduction that, in a multinomial choice model, the point $x = 0$ could arise via re-centering.

The constructive identification argument is quite simple. The $P(x) = P\left(x, F^0\right)$ are the observed choice probabilities in the data. We illustrate the argument for the special case where $K = 2$ and so $x'\beta = x_1\beta_1 + x_2\beta_2$. At $x_1 = x_2 = 0$,

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

where β_1 arises from the chain rule and the expression identifies the mean of β_1, because $P\left(x, F^0\right)$ is in the data and $g^{(1)}(0)$ is a known constant that does not depend on β. Likewise, $\left.\frac{\partial P(x, F^0)}{\partial x_2}\right|_{x=0} / g^{(1)}(0)$ equals $E\left[\beta_2\right]$, $\left.\frac{\partial^2 P(x, F^0)}{\partial x_1 \partial x_2}\right|_{x=0} / g^{(2)}(0)$ equals $E\left[\beta_1\beta_2\right]$, and $\left.\frac{\partial^2 P(x, F^0)}{\partial x_1^2}\right|_{x=0} / g^{(2)}(0)$ equals $E\left[\beta_1^2\right]$. Additional derivatives will identify the other moments of $\beta = (\beta_1, \beta_2)$. We make no assumption that the components β_1 and β_2 are independently distributed; F^0 is an unrestricted joint distribution.

Theorem 11. *Assume that the vector α^0 is identified using some auxiliary argument. Suppose Assumptions 9 and 10 hold.*

- *Suppose Assumptions 7 and 8 also hold. Then the true F^0 is identified.*
- *Assume the first L derivatives of $g(c)$ with respect to c are nonzero when evaluated at the scalar argument $c = 0$. Then all moments of β up to order L (including cross moments) are identified.*

The proof is in the Appendix. Note the approach's simplicity: we need only to check for non-zero derivatives of $g(c)$ at $c = 0$. This technique can be applied to show the identification of many differentiable economic models that use linear indices $x'\beta$. The approach is constructive: if $g^{(2)}(0) \neq 0$, we can identify all own second moments and all cross-partial moments between random coefficients. If the first 100 derivatives of $g(c)$ at $c = 0$ are nonzero, then we identify at least the first 100 moments of the random coefficients.

6. Identification of the random coefficients logit model

6.1. Homogeneous parameters for the logit

Our leading example of a mixtures model is the random coefficients logit as outlined in the introduction. We first show we can identify the homogeneous terms, α_j^0 for all j. Consider the point $x = 0$. Algebra shows that

$$\log \Pr\left(j \mid x = 0;\, F^0, \alpha^{J,0}\right) - \log \Pr\left(0 \mid x = 0;\, F^0, \alpha^{J,0}\right)$$
$$= \alpha_j^0 \quad \forall j = 1, \ldots, J,$$

where $\Pr\left(j \mid x = 0;\, F^0, \alpha^{J,0}\right)$ is identified from the data for all $j = 0, 1, \ldots, J$. The vector $\alpha^{J,0}$ is identified at the point $x = 0$.

Following this, we can show the parameters (α, γ_w') in $\alpha_j \equiv \alpha_j(w) = \alpha + w_j'\gamma_w$ are identified if α_j also depends on some covariates w_j. In this case let $\Pr(j \mid x, w) = \Pr\left(j \mid x, w; F^0, \alpha^0, \gamma_w^0\right)$ be the choice probability for product j when the regressors are x and w. The population linear regression of $\log \Pr(j \mid x = 0, w; F^0, \alpha^0, \gamma_w^0) - \log \Pr\left(0 \mid x = 0, w; F^0, \alpha^0, \gamma_w^0\right)$ on a constant and the vector w_j will identify the constant term α^0 and the homogeneous coefficients γ_w^0 as long as the density of $w = (w_1', \ldots, w_J')$ is strictly positive on its support at $x = 0$ and $E\left[(1, w_j')'(1, w_j') \mid x = 0\right]$ is nonsingular. If some elements of w_j are discrete, we use a density with respect to the counting measure for the discrete elements. Thus, the homogeneous terms are identified from differences in market shares when all products are evaluated at $x = 0$. The identification of F^0 below can proceed at any given value of $w = \tilde{w}$ in the support of w as long as $x = 0$ is allowed at $\tilde{w}$.

Remark 12. Assume that x and w jointly have product support $\mathcal{X} \times \mathcal{W}$. The identification argument for (α^0, γ_w^0) directly relies on a set of (x, w) of measure zero on $\mathcal{X} \times \mathcal{W}$ only because of the focus on the point $x_1 = \cdots = x_J = 0$, not because of any additional restrictions on w.

We collect these conditions for the identification of (α^0, γ_w^0).

Assumption 13. The density of $w = (w_1', \ldots, w_J')$ is strictly positive on its support at $x = 0$ and $E\left[(1, w_j')'(1, w_j') \mid x = 0\right]$ is nonsingular.

Below, unless necessary, we write $\alpha_j \equiv \alpha_j^0(w_j) = \alpha^0 + w_j'\gamma_w^0$ and $\alpha^J \equiv \alpha^{J,0}(w) = (\alpha_1^0(w_1), \ldots, \alpha_J^0(w_J))$ for ease of notation that covers both the cases where α_j is a constant and where α_j depends on some covariates w_j.

6.2. Distribution of the random coefficients for the logit

Using some duplication of notation, we can fit the mixed logit model into the mixtures framework by defining the logit choice probabilities for some particular choice j as

$$g_j\left(\alpha^J, x_1'\beta, \ldots, x_J'\beta\right) = \frac{\exp\left(\alpha_j + x_j'\beta\right)}{1 + \sum_{j'=1}^{J} \exp\left(\alpha_{j'} + x_{j'}'\beta\right)}.$$

Let $x_{j'} = 0$ for all $j' \neq j$. With one outside good and J inside goods, the choice probability of alternative j given β is

$$g_j\left(\alpha^J, 0, \ldots, x_j'\beta, \ldots, 0\right)$$
$$= \frac{\exp\left(x_j'\beta\right)}{\exp\left(-\alpha_j\right) + \sum_{j' \neq j} \exp\left(\alpha_{j'} - \alpha_j\right) + \exp\left(x_j'\beta\right)}.$$

Then we obtain the integrated choice probability $P_j\left(0, \ldots, x_j, \ldots, 0, F, \alpha^J\right)$

$$= \int g_j\left(\alpha^J, 0, \ldots, x_j'\beta, \ldots, 0\right) \mathrm{d}F(\beta), \qquad (6)$$

where $P_j\left(0, \ldots, x_j, \ldots, 0, F, \alpha^J\right)$ denotes the conditional choice probability of the good j at $x = (0', \ldots, x_j', \ldots, 0')$ and α^J.

Define $A_j\left(\alpha^J\right) = \exp\left(-\alpha_j\right) + \sum_{j' \neq j} \exp\left(\alpha_{j'} - \alpha_j\right)$ and, in another duplication of notation,

$$g_j(\alpha^J, c) = g_j(A_j(\alpha^J), c) = \frac{\exp(c)}{A_j\left(\alpha^J\right) + \exp(c)},$$

which is a function of a single argument c given A_j. Therefore, the formulation (6) is a special case of (3), where we take $P\left(x_j, F, \alpha\right) = P_j\left(0, \ldots, x_j, \ldots, 0, F, \alpha^J\right)$ and $g\left(\alpha^J, x_j'\beta\right) = g_j\left(\alpha^J, x_j'\beta\right)$. The choice j identification focuses on is arbitrary.

6.2.1. Non-constructive identification for the logit

Assumption 14. At some $w = \tilde{w}$, the support of x, $\mathcal{X}$, contains $x = 0$, but not necessarily an open set surrounding it. Further, the support contains a nonempty open set of points (open in $\mathbb{R}^K$) of the form $(x_1', \ldots, x_{j-1}', x_j', x_{j+1}', \ldots, x_J') = (0', \ldots, 0', x_j', 0', \ldots, 0')$.

We require that this support condition hold only at some values of w if α^J depends on w. If x is independent of w or if the support of (x, w) is a product space $\mathcal{X} \times \mathcal{W}$, the support condition does not depend on w. The identification approach needs $x = 0$ to be in the support of x so that the homogeneous parameters are identified and so that identification can exclusively focus on variation in the characteristics of choice j. The point $x = 0$ otherwise plays no role in the identification of the distribution of random coefficients. In other words, identification of the random coefficients can come from a small open set of x_j values far from $x_j = 0$. The difference between Assumptions 10 and 14 is subtle, but the role of the point $x = 0$ in identification of F^0 is quite different in the constructive and non-constructive identification approaches.

Theorem 15. Let the true model be the multinomial logit and let Assumptions 2, 13 and 14 hold. The homogeneous parameters $\alpha^{J,0}$

are identified. Also, the distribution $F^0 \in \mathcal{F}(\mathcal{B})$ is uniformly identified over choices of $(\mathcal{B}, \mathcal{X}_0^K)$, where $\mathcal{X}_0^K \subseteq \mathbb{R}^K$ are nonempty open subsets of the space of characteristics of one particular product, $\mathcal{X}^K$.

Identification holds in any nonempty open set of product characteristics satisfying the conditions in the theorem. This theorem is a specialization of Theorem 4 to the multinomial logit. The logit $g_j\left(\alpha^J, x_j'\beta\right)$ is nonzero at all $x_j \in \mathcal{X}^K$ for finite α^J, is bounded, and is real analytic. Real analyticity holds because the function $\exp\left(\cdot\right)$ is real analytic and the function $g_j(\alpha^J, c)$ is formed by the addition and division of never zero real analytic functions, and so is itself real analytic (Krantz and Parks, 2002). Thus, the distribution of random coefficients in the multinomial logit is non-constructively identified.

6.2.2. Constructive identification

Let $g_j^{(l)}\left(\alpha^J, 0\right)$ be the l-th derivative of $g_j\left(\alpha^J, c\right) = \frac{\exp(c)}{A_j(\alpha^J)+\exp(c)}$ with respect to c evaluated at $c = 0$. Define the set

$$\mathcal{A} = \left\{\alpha^J \in \mathbb{R}^J \mid g_j^{(l)}\left(\alpha^J, 0\right) \neq 0 \text{ for all integer } l \geq 1\right\}. \qquad (7)$$

This is the set of values of homogeneous terms, identified in the first stage, where the logit has nonzero derivatives and hence all moments of β are identified using Theorem 11. If $\mathcal{A} = \mathbb{R}^J$, then we would write the logit model is identified. Unfortunately, $\mathcal{A} \subset \mathbb{R}^J$, although we will show that $\mathbb{R}^J \setminus \mathcal{A}$ is a set of measure 0.

Assumption 16. At some $w = \tilde{w}$, the support $\mathcal{X}$ contains a nonempty open set of points (open in $\mathbb{R}^K$) of the form $(x_1', \ldots, x_{j-1}', x_j', x_{j+1}', \ldots, x_J') = (0', \ldots, 0', \tilde{x}_j', 0', \ldots, 0')$ surrounding $\tilde{x}_j = 0$.

Theorem 17. Let Assumptions 7, 9, 13 and 16 hold. The homogeneous parameters $\alpha^{J,0}$ are identified. Also the true F^0 is identified for any $\alpha^J \in \mathcal{A}$. Further, $\mathcal{A}$ is a set of measure 1 in $\mathbb{R}^J$.

Identification of F^0 follows directly from Theorem 11 and the definition of $\mathcal{A}$ once Assumption 8(1) is satisfied. A function that is real analytic is infinitely differentiable. The proof in the Appendix shows that the set $\mathcal{A}$ has measure 1. We note that α^J is always identified from the data at $x = 0$ and whether $\alpha^J \in \mathcal{A}^L$ can be computationally tested using computer algebra software, where

$$\mathcal{A}^L = \left\{\alpha^J \in \mathbb{R}^J \mid g_j^{(l)}\left(\alpha^J, 0\right) \neq 0 \text{ for all integer } 1 \leq l \leq L\right\}$$

and L is the maximum order of the derivative considered by the computer. As variation in w causes variation in $\alpha_j = \alpha + w_j'\gamma_w$, variation in w (allowing regressors without random coefficients) ensures that the true F^0 is identified at all possible α^J values, and not just identified for any $\alpha^J \in \mathcal{A}$.

As stated previously, our constructive identification argument uses only variation in product characteristics x_j around zero or around zero after centering, e.g., we can redefine $x_{j,k} = x_{j,k} - \bar{x}_k$ where $\bar{x}_k$ is a constant term that is the same across the inside goods. Because we do not advocate analog estimation (instead preferring our mixtures estimator in Fox et al. (forthcoming) or another mixtures estimator), we do not see using thin slices of data in identification as a problem for estimation.

7. Extensions

7.1. Identification for higher order terms and interactions

Our identification results have not been extended to models with higher order polynomial terms and interaction terms. Our support conditions rule out these models for both the non-constructive and constructive identification arguments. When $x_{j,k}$ includes higher order terms or interactions of characteristics, the support of x_j cannot include a nonempty open set because $x_{j,k}$ has components that are nonlinearly dependent. This is a limitation of our identification results.

7.2. Need for a special regressor for a random coefficient for the constant term

Consider the utility specification $u_{i,j} = \alpha_i + x_j'\beta_i + \epsilon_{i,j}$, where we now allow for a random coefficient α_i on the constant term, which affects the utility of the inside relative to the outside goods. Consider the case of $J = 1$, or one inside and one outside good. In this case, the unspecified density $f_\alpha(\alpha)$ subsumes the type I extreme value density on $\epsilon_{i,j}$. This puts us in the model of Ichimura and Thompson (1998), where all previous identification results require special regressors (regressors with large support).

7.3. Endogeneity

If there is endogeneity in price due a demand shock or omitted product characteristic ξ_j in the utility of choice j, one can adopt Kim and Petrin (2009)'s control function approach. Kim and Petrin show that under a set of conditions (including restrictions on the supply side), one can include a proxy for ξ_j that is a (to be identified) nonparametric function of the residuals for all products from first stage regressions of pricing equations. Alternatively, one can identify ξ_j in a first stage using both a special regressor and instruments with the approach of Berry and Haile (2010). Either way, ξ_j or proxies for it can be added to the characteristic vector x_j and identification using our approach can be considered.

8. Conclusions

The random coefficients logit model has been used in empirical studies for over thirty years. In contrast to other work on identification in binary and multinomial choice, we exploit the type I extreme value distribution on the additive errors to show that the distribution of random coefficients is nonparametrically identified under an alternative set of assumptions, which are non-nested with those used in other approaches. By exploiting this special structure, we eliminate assumptions about the large support and the signs of coefficients on special regressors. We do use identification at one particular point ($x = 0$), but, in the non-constructive identification theorem, only to identify homogeneous parameters such as the product intercepts α in a first stage and to focus on variation in the characteristics of only one choice j. An open set around $x = 0$ is the only source of identification for our constructive identification result. From an econometric theory perspective, either approach allows complete proofs for the consistency of nonparametric estimators of the distribution of random coefficients.

The proof of identification is also comforting to empirical researchers. Prior to our theorem, it was not known whether variation in x was sufficient to identify the distribution of the random coefficients in these models. One possibility was that the normality assumptions typically imposed on the distribution were crucial to identification: without restricting attention to a particular parametric functional form, two different distributions of random coefficients would be consistent with the data in the model, even with data on a continuum of x. We show that indeed the random coefficients logit model is nonparametrically identified, which provides a solid econometric foundation for its widespread use in empirical work. We can condition on discrete covariates, but, like the rest of the literature, we cannot point identify the distribution of random coefficients on discrete characteristics. Our identification results have not been extended to models with higher order terms or interaction terms.

Acknowledgments

Appendix

A.1. Proof of Lemma 5

We rephrase the proof of Theorem 3.8 in Stinchcombe and White (1998) in terms of our identification problem; otherwise our proof is essentially the same.

The forward direction of Lemma 5 holds because a stronger definition of identification implies a weaker definition. For the reverse direction of Lemma 5, assume to the contrary: the distribution of random coefficients is identified uniformly over $\mathcal{B}$ with a nonempty open set $\mathcal{T}_0 \subset \mathcal{T}$ but there exists a nonempty open set $\tilde{\mathcal{T}} \subset \mathcal{T}$ and a compact set $\tilde{\mathcal{B}}$ where identification fails. The lack of identification means that there exist $F^0, F^1 \in \mathcal{F}(\tilde{\mathcal{B}})$ such that $F^0 \neq F^1$ but $P(x, F^0) = P(x, F^1)$ for all $x \in \tilde{\mathcal{T}}$. It follows that $\Delta(x) \equiv \int_{\tilde{\mathcal{B}}} g(x'\beta)d\left(F^0(\beta) - F^1(\beta)\right) = 0$ for all $x \in \tilde{\mathcal{T}}$.

Because $\tilde{\mathcal{B}}$ is compact and $g(\cdot)$ is real analytic, $\Delta(x)$ is itself a real analytic function. If a real analytic function equals to 0 on an open set, it equals 0 everywhere. If $\Delta(x) = 0$ everywhere, then $F^0 \in \mathcal{F}(\tilde{\mathcal{B}})$ cannot be identified on the set $\mathcal{T}_0$, which gives a contradiction.

A.2. Proof of Lemma 6

For completeness, we reconstruct the proof of Theorem 5 in Hornik (1991) in terms of our identification problem. Most of the proof steps are essentially identical.

For the purpose of contradiction pick a $F^1 \in \mathcal{F}(\mathcal{B})$ such that $F^1 \neq F^0$ and suppose $g(\cdot)$ is bounded and nonconstant such that $\int_{\mathcal{B}} g(x'\beta)d\left(F^0(\beta) - F^1(\beta)\right) = 0$ for all $x \in \mathbb{R}^K$. Let σ denote the finite signed measure (concentrated) on $\mathcal{B}$ corresponding to $F^0 - F^1$. Then fix $\eta \in \mathbb{R}^K$ and let σ_η be the finite signed measure on $\mathbb{R}$ induced by the transformation $\beta \mapsto \eta'\beta$ in the following sense: for all Borel sets of $\mathbb{R}$ we have $\sigma_\eta(C) = \sigma\{\beta \in \mathcal{B}: \eta'\beta \in C\}$.

Then at least for all bounded functions χ on $\mathbb{R}$, $\int_{\mathcal{B}} \chi(\eta'\beta)d\left(F^0(\beta) - F^1(\beta)\right) = \int_{\mathbb{R}} \chi(t)d\sigma_\eta(t)$. Therefore by assumption on $g(\cdot)$, we have

$$0 = \int_{\mathcal{B}} g(\lambda\eta'\beta)d\left(F^0(\beta) - F^1(\beta)\right) = \int_{\mathbb{R}} g(\lambda t)d\sigma_\eta(t)$$

for all $\lambda \in \mathbb{R}$. To simplify notation, denote $L = L^1(\mathbb{R})$ for the space of integrable functions on $\mathbb{R}$ (with respect to Lebesgue measure) and $M = M(\mathbb{R})$ for the space of finite signed measures on $\mathbb{R}$. For $f \in L$, $\hat{f}$ denotes the Fourier transform. Similarly, for $\mu \in M$, $\hat{\mu}$ denotes the Fourier transform.

Because $g(0) \neq 0$ and setting $\lambda = 0$, we find that in particular

$$\int_{\mathbb{R}} d\sigma_\eta(t) = \hat{\sigma}_\eta(0) = 0. \tag{8}$$

For $\eta = 0$, σ_0 is concentrated at $t = 0$ and $\sigma_0\{0\} = \hat{\sigma}_0 = 0$, hence $\sigma_0 = 0$.

Now consider $\eta \neq 0$ and the integral

$$\int_{\mathbb{R}} g(\lambda t)d\sigma_\eta(t) = 0 \tag{9}$$

and note that σ_η is absolutely continuous with respect to Lebesgue measure (on $\mathbb{R}$) by construction of σ_η from σ. Let h be the corresponding Radon–Nikodym derivative and note $h \in L$. Then $\hat{h} = \hat{\sigma}_\eta$ and in particular from (8) we have $\hat{h}(0) = 0$. (9) is then equivalent to $\int_{\mathbb{R}} g(\lambda t)h(t)dt = 0$. Rewriting $\lambda = 1/\tau$ with $\tau \neq 0$ and applying the change of variables $t \mapsto \tau t + s$, we obtain for all nonzero real τ

$$\int_{\mathbb{R}} g\left(t + \frac{s}{\tau}\right)h(\tau t + s)dt = 0. \tag{10}$$

Write $M_\tau h(t)$ for $h(\tau t)$. The above equation implies that $\int_{\mathbb{R}} g(t + c)f(t)dt$ for some c vanishes for all f contained in the closed

translation invariant subspace I spanned by the family $M_\tau h$, $\tau \neq 0$. The subspace I is also an ideal in L by Theorem 7.1.2 in Rudin (1967). Following the notation in Rudin (1967) (also in Hornik (1991)), write $Z(f)$ for the set of all $\omega \in \mathbb{R}$ where the Fourier transform $\hat{f}(\omega)$ for $f \in L$ vanishes and define $Z(I)$, the zero set of I, as the set of ω where the Fourier transforms of all functions in I vanish.

For the purpose of contradiction, suppose that h is nonzero. As $\widehat{M_\tau h}(\omega) = \hat{h}(\omega/\tau)/\tau$ and $\hat{h}(0) = 0$, following exactly the same argument in Hornik (1991) we conclude that $Z(I) = \{0\}$ and also that I is precisely the set of all integrable functions f with $\int_{\mathbb{R}} f(t)\mathrm{d}t = \hat{f}(0) = 0$. Because I is an ideal subspace of L and h is nonzero, the statements above together with (10) imply that now the integral $\int_{\mathbb{R}} g(t+c)f(t)\mathrm{d}t$ for some c vanishes for all integrable functions $f \in L$ that have zero integral. As Hornik (1991) argues, this implies that $g(\cdot)$ must be constant, which was ruled out by our assumption that $g(\cdot)$ is bounded and nonconstant.[2] We therefore conclude $h = 0$ and thus $\hat{h} = \hat{\sigma}_\eta$ is identically zero. By the uniqueness Theorem 1.3.7(b) in Rudin (1967), we conclude $\sigma_\eta = 0$ for all $\eta \in \mathbb{R}^K$.

To complete the proof, let $\hat{\sigma}(\eta) = \int_{\mathcal{B}} \exp(i\eta'\beta)\mathrm{d}\sigma(\beta)$ be the Fourier transform of σ at η. It follows that

$$\hat{\sigma}(\eta) = \int_{\mathcal{B}} \exp(i\eta'\beta)\mathrm{d}\sigma(\beta) = \int_{\mathbb{R}} \exp(it)\mathrm{d}\sigma_\eta(t) = 0,$$

and thus $\hat{\sigma} = 0$. Again invoking the uniqueness Theorem 1.3.7(b) in Rudin (1967), we conclude $\sigma = 0$ implying $F^1 = F^0$. This completes the proof.

A.3. Proof of Theorem 11

First we introduce some notation for gradients of arbitrary order, which we need because $F(\beta)$ has a vector of K arguments, β. Let t be a vector of length T. For a function $h(t)$, we denote the $1 \times K^\upsilon$ block vector of υ-th order derivatives as $\nabla^\upsilon h(t)$. $\nabla^\upsilon h(t)$ is defined recursively so that the k-th block of $\nabla^\upsilon h(t)$ is the $1 \times T$ vector $h_k^\upsilon(t) = \partial h_k^{\upsilon-1}(\theta)/\partial t'$, where $h_k^{\upsilon-1}$ is the k-th element of $\nabla^{\upsilon-1} h(t)$. Using a Kronecker product $\otimes$, we can write $\nabla^\upsilon h(t) =$

$$\underbrace{\frac{\partial^\upsilon h(t)}{\partial t' \otimes \partial t' \otimes \cdots \otimes \partial t'}}_{\upsilon \text{ Kronecker product of } \partial t'}.$$

Take the derivatives with respect to the covariates x on both sides of $P(x, F) = \int g(x'\beta)\mathrm{d}F(\beta)$ and evaluate the derivatives at $x = 0$. By Assumption 8, for any $\upsilon = 1, 2, \dots$ and the chain rule repeatedly applied to the linear index $x'\beta$,

$$\nabla^\upsilon P(x, F)|_{x=0} = \int g^{(\upsilon)}(x'\beta)|_{x=0} \{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\} \mathrm{d}F(\beta)$$

$$= g^{(\upsilon)}(0) \int \{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\} \mathrm{d}F(\beta). \tag{12}$$

---

[2] If $g(\cdot)$ is not constant, then we can easily construct an example such that $\int_{\mathbb{R}} g(t+c)f(t)\mathrm{d}t \neq 0$ when $\int_{\mathbb{R}} f(t)\mathrm{d}t = 0$. For an arbitrary constant δ, let $E_1 = \{t \in \mathbb{R} | g(t+c) > \delta\}$ and $E_2 = \{t \in \mathbb{R} | g(t+c) \leq \delta\}$ such that the (Lebesgue) measures on E_1 and E_2 are both positive (so this rules out $g(\cdot)$ is a constant function), respectively as $\mu_1 = \mu\{E_1\}$ and $\mu_2 = \mu\{E_2\}$. Define $f(t) = (1\{t \in E_1\}/\mu_1 - 1\{t \in E_2\}/\mu_2)f_\mu(t)$ with $f_\mu(t)$ the pdf with respect to the measure μ and note by construction $\int_{\mathbb{R}} f(t)\mathrm{d}t = 0$. However, consider

$$\int_{\mathbb{R}} g(t+c)f(t)\mathrm{d}t = \int_{\mathbb{R}} \{g(t+c) - \delta\}f(t)\mathrm{d}t$$

$$= \int_{\mathbb{R}} \{g(t+c) - \delta\}1\{t \in E_1\}f_\mu(t)\mathrm{d}t/\mu_1$$

$$+ \int_{\mathbb{R}} \{\delta - g(t+c)\}1\{t \in E_2\}f_\mu(t)\mathrm{d}t/\mu_2 > 0, \tag{11}$$

where the first equality holds because $\int_{\mathbb{R}} \delta f(t)\mathrm{d}t = 0$ and the last inequality holds because both integrals in (11) are nonnegative and at least one is strictly positive, noting that $g(t+c) - \delta > 0$ for $t \in E_1$ and $\delta - g(t+c) \geq 0$ for $t \in E_2$.

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 $\upsilon = 1, 2, \dots$, we obtain the υ-th moments of β for all $\upsilon \geq 1$. Now by Assumption 7, F satisfies the Carleman condition. Therefore, F^0 is identified since a probability measure satisfying the Carleman condition is uniquely determined by its moments.

A.4. Proof of Theorem 17

Identification arises from identifying all moments, as in Theorem 11. We wish to show that the set $\mathcal{A}$ as defined in (7) has measure 1 in $\mathbb{R}^J$.

Let D_c be the derivative operator with respect to c. We suppress $A_j(\alpha^J)$'s dependence on α^J and write $A_j = A_j(\alpha^J)$. For this purpose, we first obtain the derivatives of $g_j(\alpha^J, c)$ with respect to c,

$$D_c g_j(\alpha^J, c) = (A_j + e^c)^{-2} A_j e^c,$$

$$D_c^2 g_j(\alpha^J, c) = (A_j + e^c)^{-3} (A_j^2 e^c - A_j e^{2c})$$

$$D_c^3 g_j(\alpha^J, c) = (A_j + e^c)^{-4} (A_j^3 e^c - 4A_j^2 e^{2c} + A_j e^{3c}), \dots.$$

For $p \geq 3$, now we write the $(p-1)$-th derivative as $D_c^{p-1} g_j(\alpha^J, c) = (A_j + e^c)^{-p} \sum_{l=1}^{p-1} \gamma_{p-l}^{(p)} A_j^{p-l} e^{lc}$. Then, we can write the p-th derivative as

$$D_c^p g_j(\alpha^J, c) = \left[\frac{1}{(A_j + e^c)^{p+1}} \sum_{l=1}^p \gamma_{p+1-l}^{(p+1)} A_j^{p+1-l} e^{lc} \right] \tag{13}$$

$$= D_c D_c^{p-1} g_j(\alpha^J, c) = D_c \left[\frac{1}{(A_j + e^c)^p} \sum_{l=1}^{p-1} \gamma_{p-l}^{(p)} A_j^{p-l} e^{lc} \right]$$

$$= \frac{1}{(A_j + e^c)^p} \sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p-l} e^{lc} - \frac{1}{(A_j + e^c)^{p+1}} p \sum_{l=1}^{p-1} \gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c}$$

$$= \frac{1}{(A_j + e^c)^{p+1}} \left((A_j + e^c) \sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p-l} e^{lc} - \sum_{l=1}^{p-1} p\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} \right)$$

$$= \frac{1}{(A_j + e^c)^{p+1}} \left(\sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p+1-l} e^{la} + \sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} \right.$$
$$\left. - \sum_{l=1}^{p-1} p\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} \right)$$

$$= \frac{1}{(A_j + e^c)^{p+1}} \left(\begin{array}{l} \gamma_{p-1}^{(p)} A_j^p e^c + \sum_{l'=2}^{p-1} l'\gamma_{p-l'}^{(p)} A_j^{p+1-l'} e^{l'c} \\ + \sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} - \sum_{l=1}^{p-1} p\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} \end{array} \right) \tag{14}$$

$$= \frac{1}{(A_j + e^c)^{p+1}} \left(\begin{array}{l} \gamma_{p-1}^{(p)} A_j^p e^c + \sum_{l=1}^{p-2} (l+1)\gamma_{p-l-1}^{(p)} A_j^{p-l} e^{(l+1)c} \\ + \sum_{l=1}^{p-1} l\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} - \sum_{l=1}^{p-1} p\gamma_{p-l}^{(p)} A_j^{p-l} e^{(l+1)c} \end{array} \right) \tag{15}$$

$$= \frac{1}{(A_j + e^c)^{p+1}} \left(\begin{array}{l} \gamma_{p-1}^{(p)} A_j^p e^c - \gamma_1^{(p)} A_j^1 e^{pc} \\ + \sum_{l=1}^{p-2} \left\{ (l+1)\gamma_{p-l-1}^{(p)} + l\gamma_{p-l}^{(p)} - p\gamma_{p-l}^{(p)} \right\} A_j^{p-l} e^{(l+1)c} \end{array} \right), \tag{16}$$

where in (14) and (15), we take out the first element in the first sum and change the index l' to $l+1$. (16) is obtained by rearranging terms and collecting coefficients on $A_j^{p-l} e^{(l+1)c}$ for $j = 1$ to $p - 2$.

To fix the undetermined coefficients $\gamma_{p-l}^{(p)}$'s, we compare the coefficients from (13) and (16) and obtain

$$\sum_{l=1}^p \gamma_{p+1-l}^{(p+1)} A_j^{p+1-l} e^{lc}$$

$$= \gamma_{p-1}^{(p)} A_j^p e^c + \sum_{l=1}^{p-2} \left\{ (l+1)\gamma_{p-l-1}^{(p)} + l\gamma_{p-l}^{(p)} - p\gamma_{p-l}^{(p)} \right\}$$

$$\times A_j^{p-l} e^{(l+1)c} - \gamma_1^{(p)} A_j^1 e^{pc}.$$

We find

$$\gamma_p^{(p+1)} = \gamma_{p-1}^{(p)} \tag{17}$$

$$\gamma_{p-l}^{(p+1)} = (l+1)\,\gamma_{p-l-1}^{(p)} - (p-l)\,\gamma_{p-l}^{(p)} \quad \text{for } p \geq 3 \tag{18}$$

$$\gamma_1^{(p+1)} = -\gamma_1^{(p)}. \tag{19}$$

This system generates the coefficients for all $p \geq 1$. For the initial value, we obtain $\gamma_1^{(2)} = 1$. When $p = 2$, we find

$$\gamma_2^{(3)} = \gamma_1^{(2)} = 1, \qquad \gamma_1^{(3)} = -\gamma_1^{(2)} = -1$$

and when $p = 3$, we find

$$\gamma_3^{(4)} = \gamma_2^{(3)} = 1, \qquad \gamma_2^{(4)} = 2\gamma_1^{(3)} - 2\gamma_2^{(3)} = -4,$$

$$\gamma_1^{(4)} = -\gamma_1^{(3)} = 1.$$

Now we examine whether $D_c^p g_j\left(\alpha^J, c\right)\big|_{c=0}$ can take the value of zero for some A_j (and hence for some α^J) at some p. For this purpose, we evaluate the derivatives at $c = 0$ and obtain expressions with respect to A_j for the p-th order derivative as $D_c^p g_j\left(\alpha^J, c\right)\big|_{c=0} = \frac{1}{(A_j+1)^{p+1}} \sum_{l=1}^{p} \gamma_{p+1-l}^{(p+1)} A_j^{p+1-l} = 0$ for $p \geq 1$. This is equivalent to solving

$$\sum_{l=1}^{p} \gamma_{p+1-l}^{(p+1)} A_j^{p-l} = 0. \tag{20}$$

We note that, when α_j's are not equal to zero, we have some non-integer solutions of the Eq. (20). Below we, however, show that the set

$$\mathcal{A}^C = \left\{ \alpha^J \mid D_c^p g_j\left(\alpha^J, c\right)\big|_{c=0} = 0 \right.$$
$$\left. \text{for at least one } p \geq 1, \alpha^J \in \mathbb{R}^J \right\}$$

has measure zero in $\mathbb{R}^J$ and this will complete our proof.

The set of values $\mathbb{A} \subset \mathbb{R}_+$ that collect values of A_j that solve the Eq. (20) for at least one p is countable because for any order p the Eq. (20) has at most p number of solutions, i.e., we can find an injective function that maps $\mathbb{A}$ to $\mathbb{N}$. Now consider any element $\tilde{A} \in \mathbb{A}$. We claim that the set of values $\tilde{\mathcal{A}}(\tilde{A}) = \{ \alpha^J \mid A_j(\alpha^J) = \tilde{A}, \alpha^J \in \mathbb{R}^J \}$ has measure zero in $\mathbb{R}^J$ because $\tilde{\mathcal{A}}$ is at most a subset of the vector space of $\mathbb{R}^{J-1}$. By construction, we have $\mathcal{A}^C = \cup_{\tilde{A} \in \mathbb{A}} \tilde{\mathcal{A}}(\tilde{A}) = \cup_{\tilde{A} \in \mathbb{A}} \{ \alpha^J \mid A_j(\alpha^J) = \tilde{A}, \alpha^J \in \mathbb{R}^J \}$. Finally we conclude that the set $\mathcal{A}^C$ has measure zero in $\mathbb{R}^J$ because a countable union of measure zero sets has measure zero, as $\tilde{\mathcal{A}}(\tilde{A})$ has measure zero for all $\tilde{A} \in \mathbb{A}$ and the set $\mathbb{A}$ is countable. This completes the proof.

Remark 18. Although it is not necessary, we can simplify our proof when $\alpha_j = 0$ for all j, i.e., all goods are symmetric. The condition of nonzero derivatives is obtained using the rational zero test. In this case we have $A_j = J$ and note that the coefficient on A_j^{p-1} (the highest order term in the equation) in (20) is equal to $\gamma_p^{(p+1)} = 1$ for all p. Also note that the constant term (the coefficient on A_j^0 in (20)), $\gamma_1^{(p+1)}$, is equal to 1 when p is odd and is equal to -1 when p is even. By the rational zero test, this implies that the only possible positive rational number solution in (20), if any, is $J = 1$ for all $p \geq 1$, so obviously the set $\mathcal{A}^C$ has measure zero (in $\mathbb{R}^J$) or is empty if $J \geq 2$.

References

Andrews, D.W.K, Schafgans, M.M.A., 1998. Semiparametric estimation of the intercept of a sample selection model. Review of Economic Studies 65 (3), 497–517.

Berry, S., Haile, P., 2010, Nonparametric identification of multinomial choice models with heterogeneous consumers and endogeneity, Yale University Working Paper.

Boyd, J.H., Mellman, R.E., 1980. Effect of fuel economy standards on the US automotive market: an hedonic demand analysis. Transportation Research B 14 (5), 367–378.

Briesch, Richard A., Chintagunta, Pradeep K., Matzkin, Rosa L., 2010. Nonparametric discrete choice models with unobserved heterogeneity. Journal of Business and Economic Statistics 28 (2), 291–307.

Burda, Martin, Harding, Matthew, Hausman, Jerry, 2008. A Bayesian mixed logit–probit for multinomial choice. Journal of Econometrics 147 (2), 232–246.

Cardell, N.S., Dunbar, F.C., 1980. Measuring the societal impacts of automobile downsizing. Transportation Research B 14 (5), 423–434.

Chiappori, P.A., Komunjer, I., 2009, On the nonparametric identification of multiple choice models, Columbia University Working Paper.

Cramér, H., Wold, H., 1936. Some theorems on distribution functions. Journal of the London Mathematical Society s1-11 (4), 290–294.

Domencich, Thomas A., McFadden, Daniel L., 1975. Urban Travel Demand: A Behavioral Analysis: A Charles River Associates Research Study. North Holland.

Fox, Jeremy T., Gandhi, Amit, 2010, Nonparametric identification and estimation of random coefficients in nonlinear economic models, University of Michigan Working Paper.

Fox, J.T., Kim, K., 2011, A simple nonparametric approach to estimating the distribution of random coefficients in structural models, University of Michigan Working Paper.

Fox, J.T., Kim, K., Ryan, S., Bajari, P., 2011, A simple estimator for the distribution of random coefficients, Quantitative Economics (forthcoming).

Gautier, Eric, Kitamura, Yuichi, 2009, Nonparametric estimation in random coefficients binary choice models, Cowles Foundation Discussion Paper No. 1721.

Hausman, Jerry, Wise, David A., 1978. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. Econometrica 46 (2), 403–426.

Heckman, J.J., Willis, R.J., 1977. A beta-logistic model for the analysis of sequential labor force participation by married women. Journal of Political Economy 85 (1), 27–58.

Hornik, Kurt, 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4, 251–257.

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

Kahn, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.

Kim, Kyoo il, Petrin, Amil, 2009, Control function corrections for unobserved factors in differentiated product models, University of Minnesota Working Paper.

Krantz, Steven G., Parks, Harold R., 2002. A Primer on Real Analytic Functions, second ed. Birkhauser.

Lewbel, Arthur, 2000. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. Journal of Econometrics 97 (1), 145–177.

Magnac, T., Maurin, E., 2007. Identification and information in monotone binary models. Journal of Econometrics 139 (1), 76–104.

Matzkin, Rosa L., 2007. Heterogeneous choice. In: Advances in Economics and Econometrics: Theory and Applications (Ninth World Congress–Volume III). (Chapter 4).

McFadden, Daniel, Train, Kenneth, 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15 (5), 447–470.

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

Rudin, W., 1967. Fourier Analysis on Groups. In: Interscience Tracts in Pure and Applied Mathematics, vol. 12. Interscience Publishers, New York.

Shohat, J.A., Tamarkin, J.D., 1943. The Problem of Moments. American Mathematics Society, Providence, RI.

Stinchcombe, Maxwell B., White, Halbert, 1998. Consistent specification testing with nuisance parameters present only under the alternative. Econometric Theory 14, 295–325.

Teicher, Henry, 1963. Identifiability of finite mixtures. Annals of Mathematical Statistics 34, 1265–1269.

Train, K.E., 2008. EM algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling 1 (1), 40–69.