The Random Coefficients Logit Model is Identified

Patrick Bajari
University of Minnesota & NBER

Jeremy T. Fox
University of Chicago & NBER

Kyoo il Kim
University of Minnesota

Stephen P. Ryan
MIT & NBER*

First Version: October 2006
This version: November 2009

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 this model is nonparametrically identified. Our approach requires variation in product characteristics in an open set and does not rely on monotonicity assumptions or so-called special regressors used in related papers. Our identification argument is constructive and may be applied to other choice models with random coefficients.

*Fox thanks the National Science Foundation, the Olin Foundation, and the Stigler Center and Bajari thanks the National Science Foundation, various grants, for generous funding.

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 models choice between one of a finite number of competing alternatives. 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, agents $i = 1, ..., N$ can choose between $j = 1, ..., J$ mutually exclusive inside goods and one outside good. The exogenous variables for choice j are in the $K \times 1$ vector $x_{i,j}$. In the example of demand estimation, $x_{i,j}$ might include the product characteristics, the price of good j and the interactions of product characteristics with the demographics of agent i. In this paper, we let $x_{i,j}$ have mean (across observations i) equal to the 0 vector, which can occur by centering each element of $x_{i,j}$ around its population mean.[1] We shall let $x_i = (x'_{i,1}, ..., x'_{i,J})$ denote the stacked vector of the all the $x_{i,j}$. 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 intercept for each choice j that we denote as α_j. Let α be the vector of the J α's. Agent i's utility for choice j is equal to

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

The outside good has a utility of $u_{i,0} = \epsilon_{i,0}$.[2] 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 CDF of $\exp(-\exp(-\epsilon_{i,j}))$. The random coefficients logit arises when β_i varies across the population, with unknown density $f(\beta_i)$. The unknown objects of estimation are the density $f(\beta_i)$ and the homogeneous intercepts α.[3] Under the standard assumption that β_i is independent of $x_{i,j}$ (we discuss endogeneity below), utility maximization leads to choice probabilities of

$$\Pr(j \mid x_i; f, \alpha) = \int \frac{\exp(\alpha_j + x_{i,j}\beta_i)}{1 + \sum_{j'=1}^{J} \exp(\alpha_{j'} + x_{i,j'}\beta_i)} f(\beta_i)\, d\beta_i. \tag{2}$$

[1] See Appendix A for details on centering.

[2] The Type I extreme value distribution gives the scale normalization for utility values. The outside good's utility gives the location normalization.

[3] If the intercepts α_j are instead heterogeneous with a nonparametric distribution, the heterogeneous $\alpha_{i,j}$ would subsume the logit errors $\epsilon_{i,j}$.

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.[4]

In this paper, we prove that the distribution $f(\beta)$ is nonparametrically identified, in the sense that the true $\left(f^0, \alpha^0\right)$ are the only the pair (f, α) that solves (2) for all j and x_i. Our approach does not rely on *a priori* sign restrictions or support conditions on the regressors, as in much of the previous and contemporary literature in this area. Our identification theorem is constructive. We first show how to recover the intercepts, α, in (2). We then demonstrate how to iteratively find all moments of β, which is sufficient to identify the density f^0 within the class of densities that are uniquely determined by all of their moments. This class is the set of probability density functions that satisfy Carleman's condition, which we review below. Our proof strategy is not unique to the logit: it could be applied to identify the density of heterogeneity in many differentiable economic models where covariates enter as linear indices. We outline the main theorem using generic notation and verify its main condition for the multinomial logit model.

While our approach 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 note, 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. We introduce a computationally simple, nonparametric sieve estimator for f^0 in Bajari, Fox, Kim and Ryan (2009) for general mixtures models. We prove consistency of our estimator for the cumulative distribution function of β in the Lévy-Prokhorov metric under the maintained assumption that the model is identified.[5] This identification theorem therefore completes our proof of consistency for the estimator of the random coefficients logit in Bajari, Fox, Kim and Ryan.[6]

[4]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 $x_{i,j}$ in the random coefficients estimation as a function of some smaller set of underlying true product characteristics. McFadden and Train do not study identification.

[5]Alternative nonparametric estimators include the Bayesian MCMC estimators in Rossi, Allenby, and McCulloch (2005) as well as Burda, Harding and Hausman (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 estimates of (f, α) in these other procedures.

[6]Strictly speaking, the consistency Theorem 3.1 in Bajari, Fox, Kim and Ryan (2009) applies to models without

The proof of identification is comforting to empirical researchers. Prior to our theorem, it was not known whether variation in x_i was sufficient to identify the pair $\left(f^0, \alpha^0\right)$. 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 related literature, Section 3 states the general result, Section 4 shows how the results apply to the random coefficients logit, and Section 5 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.

Previously, Ichimura and Thompson (1997) 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 to demand analysis, which study markets with two or more inside goods. Ichimura and Thompson identify the cumulative distribution function (CDF) of, in our notation, $(\beta, \epsilon_{i,1} - \epsilon_{i,0})$. They use a theorem due to Cramer and Wold (1934) 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 $x_{i,1}$. Similar assumptions will appear in many of the papers below. We will refer to a regressor whose random coefficient has a sign restriction and that has full support as a "special regressor." Gautier and Kitamura (2008) 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 (1997) to the case of multinomial choice.

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.

---

homogeneous parameters, such as the intercepts α. However, in Section 2.6 of that paper we discuss an extension to nonlinear least squares where location and scale parameters (homogeneous parameters) are also estimated. A previous draft of Bajari et al included the parametric verification of the regularity conditions for nonlinear least squares for models with both homogeneous parameters α and a distribution of heterogeneous parameters β.

Briesch, Chintagunta and Matzkin (2009) study the identification of a discrete choice model where the payoff to choice j is $V(j, z_j, s_i, \omega_i) - r_j + \epsilon_{i,j}$, 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. Briesch et al do not examine whether their model nests the random coefficients logit with an added special regressor: a sign restriction and large support.

Subsequent to the circulation of this theorem, Berry and Haile (2008) and Fox and Gandhi (2009) introduced identification arguments for multinomial choice models without the Type I extreme value distribution or additive errors. Like the analysis of binary choice in Ichimura and Thompson (1997), 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 full identification) a full support assumption on the corresponding k-th component $x_{k,i,j}$, for all choices $j \in J$. Berry and Haile identify the conditional-on-x_i distribution of utility values $G(u_{i,0}, u_{i,1}, \ldots, u_{i,J} \mid x_i)$ 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_i^1 and x_i^2, or $H\left(\Delta_i u \mid x_i^1, x_i^2\right)$, where

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

where $u_{i,j}\left(x_i^1\right)$ is just the realized utility value (1) for $x_i^1 = \left(x_{i,1}^1, \ldots, x_{i,J}^1\right)$. 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_i, as in $D(u_{i,1}(x), \ldots, u_{i,J}(x))$, where $u_{i,1}(x)$ 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 $\epsilon_{i,j}$, rather than trying to recover the distribution of $\epsilon_{i,j}$. This corresponds to empirical practice: the random coefficients logit is a very popular specification in applied work. Our results contribute to the literature 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. Our proof is constructive and may be applied to other choice models with random coefficients other than the random coefficients logit as we describe below.[7]

[7] In subsequent and at present in-progress work, Chiappori and Komunjer (2009) present preliminary results for achieving the same weakening of conditions on special regressors without parametric assumptions in the multinomial

Our paper, like many of the papers in the literature, focuses on continuous covariates in x. All 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.

3 General Result

When stating the general result, we shall consider an abstract model that includes the random coefficients logit without intercepts α_j as a special case. In the next section, we verify the key identification condition for the multinomial logit and allow for intercepts.

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, say $y < \frac{1}{2}$, happened or did not happen. $P(x)$ is the probability of the event happening. x is independent of β. Let $g(x, \beta)$ be the probability of an agent with characteristics β taking the action. In our framework, the researcher specifies $g(x, \beta)$. Our goal is to identify the density function $f(\beta)$ in the equation

$$P(x) = \int g(x, \beta) f(\beta) d\beta. \tag{3}$$

Identification means that a unique $f(\beta)$ solves this equation for all x. This is the definition of identification used in the statistics literature (Teicher 1963).[8]

Here we propose to identify the density of $f(\beta)$ by finding its moments when g is differentiable and satisfies the single-index condition $g(x, \beta) = g(x'\beta)$, so that $g(\cdot)$ is now a function with a single argument. 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 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 $x'\beta$.

choice model.

[8]For the consistency of a nonparametric estimator such as Bajari et al (2009), one typically needs a stronger notion of identification with positive probability: there exists an open set X in the support of x with positive probability where for any f and $\tilde{f}$, $\tilde{f} \neq f$, $\int g(x, \beta) f(\beta) d\beta \neq \int g(x, \beta) \tilde{f}(\beta) d\beta$ for all $x \in X$. Because we identify the unique f that solves (3), we know there exists at least one point $x^\star$ where $\int g(x^\star, \beta) f(\beta) d\beta \neq \int g(x^\star, \beta) \tilde{f}(\beta) d\beta$ for any $\tilde{f} \neq f$. We restrict attention to differentiable $g(x, \beta)$ in this paper. Thus, continuity of $g(x, \beta)$ extends the property $\int g(x, \beta) f(\beta) d\beta \neq \int g(x, \beta) \tilde{f}(\beta) d\beta$ to points x in an open set X around the initial point $x^\star$.

Assumption 3.1

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:* $\Sigma_{l\geq 1} m_l^{-1/l} = \infty$.

The Carleman condition gives uniqueness for distributions with unrestricted support. If the support of β is known and compact, uniqueness follows without the Carleman condition.

Assumption 3.2

- $g^{(l)}(0)$ *is nonzero and finite for all* $l \geq 1$ *where* $g^{(l)}(\cdot)$ *denotes the* l*-th derivative of* $g(\cdot)$.
- $g(t) \in C^{\infty}$ *(infinitely continuously differentiable) in a neighborhood of* $t = 0$.

Assumption 3.2 restricts the class of $g(x'\beta)$. Some classes of functions satisfy the condition but others do not. For example, if $g(w) = C \cdot \exp(w)$, then Assumption 3.2 is trivially satisfied because $g^{(l)}(0) = C$ 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 identification argument is quite simple. We illustrate the argument for the special case where $K = 2$ and so $x'\beta = x_1\beta_1 + x_2\beta_2$. At $x_1 = x_2 = 0$,

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

where β_1 arises from the chain rule and the expression identifies the mean of β_1, because $P(x)$ 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 $E[\beta_1^2]$. Additional derivatives will identify the other moments of $\beta = (\beta_1, \beta_2)$. Note that we make no assumption that the components β_1 and β_2 are independently distributed; f is an unrestricted joint density.

Theorem 3.1 *Let all elements of* x *be continuous. Assume that the support of the independent variables includes an open set* X *containing* $x = 0$.

- *Suppose Assumptions 3.1 and 3.2 hold. Then the true* f^0 *is identified.*
- *Assume the first* L *derivatives of* $g(z)$ *are nonzero when evaluated evaluated at the scalar argument* $z = 0$*. Then all moments of* β *up to order* L *(including cross moments) are identified.*

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

Even when only the first $L < \infty$ moments are identified, there is some hope in being able to identify the distribution of random coefficients. The problem of identifying a distribution uniquely from the first L moments of the corresponding random variable is known as the determinacy (unique solution) of the truncated Hamburger moment problem (Akhiezer 1965, Krein and Nudel'man 1973). Truncated moment problems are a well studied topic in probability theory. A key tool is a Hankel matrix, which is formed from the first L moments. If the Hankel matrix has a zero determinant, a unique distribution has these particular L moments. If L is finite, this distribution will have finite support. Extensions of these results exist for multidimensional random variables (Akhiezer 1965). One can consult these results to learn when knowledge of the first L moments may be sufficient for identification. What is important here is that results on the truncated moments problem exist and are not related to the type of economic model in which the unobserved heterogeneity enters.

4 Identification of the Random Coefficients Logit Model

Our leading example of a mixtures model is the random coefficients logit. The main version of the model is the one outlined in the introduction. Here we drop the consumer i subscripts for conciseness. In the logit, identification will be based on local variation around the point $x_1 = \ldots = x_J = 0$. Given our centering of the product characteristics, this is variation in the product characteristics of each product j around its mean product characteristics. This does not correspond to a case of equal non-centered product characteristics across products.

We first show we can identify the homogeneous parameters, the brand intercepts α_j. Consider the point $x = 0$. Algebra shows that

$$\log \Pr(j \mid 0;\, f, \alpha) - \log \Pr(0 \mid 0;\, f, \alpha) = \alpha_j \,\forall\, j = 1, \ldots, J$$

and for any choice of f. Thus, the homogeneous brand intercepts are identified from differences in market shares when all products are evaluated at their means. The parameters in α are treated as known parameters in what follows.

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

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

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

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

Define $A_j\left(\alpha\right)=\exp\left(-\alpha_j\right)+\sum_{j'\neq j}\exp\left(\alpha_{j'}-\alpha_j\right)$ and, in another duplication of notation, $g_j(c;\,\alpha)=\frac{e^c}{A_j(\alpha)+e^c}$, which is a function of a single argument. Also, let D_c denote the derivative operator with respect to c.

We define the set

$$\mathcal{A}=\left\{\alpha\mid D_c^p g_j(c;\,\alpha)\big|_{c=0}\neq 0\text{ for all }p\geq 1\right\}. \tag{4}$$

This is the set of brand intercepts, identified in the first stage, where the logit has nonzero derivatives and hence all moments of β are identified using Theorem 3.1. If $\mathcal{A}=\mathbb{R}^J$, then we would write the logit model with brand intercepts is identified. Unfortunately, $\mathcal{A}\subset\mathbb{R}^J$, although we will show that $\mathbb{R}^J\backslash\mathcal{A}$ is a set of measure 0.

To highlight our main result, we state as a theorem that the logit model is identified for all brand intercepts in a set of measure 1.

Theorem 4.1

- *Let each x_j for $j=1,\ldots,J$ be centered around its mean.*
- *Let all elements of the vector $x=(x_1,\ldots,x_J)$ be continuous with product supports. Let $x=0$ be in the support.*
- *Let there be $J\geq 2$ inside goods.*
- *Let Assumption 3.1 hold.*

Then the true f^0 is identified for any $\alpha\in\mathcal{A}$. Further, $\mathcal{A}$ is a set of measure 1 in $\mathbb{R}^J$.

Identification of f^0 follows directly from Theorem 3.1 and the definition of $\mathcal{A}$. The proof focuses on showing that $\mathcal{A}$ is a set of measure 1 in $\mathbb{R}^J$. We note that α is always identified from the data and whether $\alpha \in \mathcal{A}^P$ can be computationally tested using computer algebra software, where

$$\mathcal{A}^P = \left\{ \alpha \mid D_c^p g_j(c;\, \alpha)\big|_{c=0} \neq 0 \text{ for all } 1 \leq p \leq P \right\}$$

and P is the maximum order of the derivative considered by the computer.

Our identification argument uses variation in product characteristics $x_{i,j}$ around their mean, or around 0 after the centering. The centering can be done for each j separately as Appendix A shows, so the $x = 0$ point is not necessarily a point in the data with identical product characteristics for all inside goods. Because we do not advocate analog estimation (instead preferring our mixtures estimator in Bajari et al (2009) or another mixtures estimator), we do not see using thin slices of data in identification as a problem for estimation. It is possible that the means of the product characteristics are not in the support of the data generating process (say if the support is not a connected set). In that case, the centering of $x_{i,j}$ could be done at some point other than the mean and identification will still hold.

If there is endogeneity in price due a demand shock or omitted product characteristic $\xi_{i,1}, \ldots, \xi_{i,J}$, one can adopt the Kim, Petrin and Train (2009) control function approach. Say price $p_{i,j}$ is in $x_{i,j}$ In a "zeroth" stage, one can regress price $p_{i,j}$ on exogenous demand and supply shifters, together collected in the vector $z_{i,j}$. The residual from the population regression of $p_{i,j}$ on $z_{i,j}$ is the omitted product characteristic $\xi_{i,j}$, hence the omitted product attributes are identified. This residual estimate of $\xi_{i,j}$ can then be included in the vector $x_{i,j}$, and the identification arguments based on Theorem 4.1 can proceed. As $\xi_{i,j}$ is observed, $x_{i,j}$ can include interactions of $\xi_{i,j}$ with price and other observed product and consumer characteristics in $x_{i,j}$. Theorem 4.1 identifies a separate random coefficient (in the vector β) for each term involving $\xi_{i,j}$, including its interactions.

The control function approach of Kim, Petrin and Train (2009) may impose limitations on the supply side of a market compared to some other approaches in the literature, such as Berry and Haile (2008) and Chiappori and Komunjer (2009). Either way, most empirical studies using logit methods, particularly those in say environmental economics, transportation economics and marketing, do not instrument for price or other product characteristics. These studies often exploit experimental data, data like to a student's decision enroll in one of several class sessions where endogenous prices are not present, or data with high frequency price variation that is plausibly independent of unobservable demand shocks. The logit is used in many choice situations that are not about estimating the demand curve for a consumer product as a function of an endogenous price.

5 Conclusions

The random coefficients logit model has been used in empirical studies for over thirty years. In contrast to other work in the area, we exploit the type I extreme value distribution on the additive errors to show that the density of random coefficients is nonparametrically identified. By exploiting this special structure, we eliminate assumptions about the support and signs of coefficients on "special regressors." From an econometric theory perspective, this allows complete proofs for the consistency of nonparametric estimators of the density 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_i was sufficient to identify the pair $\left(f^0, \alpha^0\right)$. One possibility was that the normality assumptions typically imposed on f^0 were crucial to identification: without restricting attention to a particular parametric functional form, two f's would indeed solve (2) for all x_i, even with data on a continuum of x_i. 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.

A Centering Product Characteristics in the Random Coefficients Logit Model

We investigate to what extent normalizing the product characteristics of each product j to have a zero mean across consumers is innocuous. In the linear regression model

$$y_i = \tilde{\alpha} + \beta \tilde{x}_i + \epsilon_i,$$

one can rewrite the model as

$$y_i = \left(\tilde{\alpha} + \beta E\left[\tilde{x}_i\right]\right) + \beta\left(\tilde{x}_i - E\left[\tilde{x}_i\right]\right) + \epsilon_i = \alpha + \beta x_i + \epsilon_i,$$

where $\alpha = \tilde{\alpha} + \beta E\left[\tilde{x}_i\right]$ and $x_i = \tilde{x}_i - E\left[\tilde{x}_i\right]$. In this model, centering covariates, so that $E\left[x_i\right] = E\left[\tilde{x}_i - E\left[\tilde{x}_i\right]\right] = 0$, just changes the interpretation of the intercept, α.

Now consider the linear regression model with random coefficients, or

$$\begin{aligned} y_i = \tilde{\alpha} + \beta_i \tilde{x}_i + \tilde{\epsilon}_i = & \\ & \left(\tilde{\alpha} + E\left[\beta_i\right] E\left[\tilde{x}_i\right]\right) + \beta_i\left(\tilde{x}_i - E\left[\tilde{x}_i\right]\right) + \left(\tilde{\epsilon}_i + \left(\beta_i - E\left[\beta_i\right]\right) E\left[\tilde{x}_i\right]\right) = \\ & \alpha + \beta_i x_i + \epsilon_i. \end{aligned} \quad (5)$$

Now the interpretation of α, x and ϵ have changed. In particular, $\epsilon_i = \tilde{\epsilon}_i + \left(\beta_i - E\left[\beta_i\right]\right) E\left[\tilde{x}_i\right]$. The model before the centering is behaviorally equivalent to the model after the centering.

The model this paper focuses on is the random coefficients logit, in which

$$u_{i,j} = \tilde{\alpha}_j + \beta_i \tilde{x}_{i,j} + \tilde{\epsilon}_{i,j}$$

and the distribution of $\tilde{\epsilon}_{i,j}$ has the type I extreme value distribution. The change of variables in (5) is strictly speaking not permissible, because $\epsilon_{i,j} = \tilde{\epsilon}_{i,j} + \left(\beta_i - E\left[\beta_i\right]\right) E\left[\tilde{x}_{i,j}\right]$ may not have the type I extreme value when $\tilde{\epsilon}_{i,j}$ has that distribution. The parametric assumption on $\tilde{\epsilon}_{i,j}$ complicates changing the location of covariates, unlike models where the distribution of $\epsilon_{i,j}$ is unspecified.

If there are no outside goods or the outside good has measured product characteristics $\tilde{x}_{i,0}$, then the term $\left(\beta_i - E\left[\beta_i\right]\right) E\left[\tilde{x}_{i,j}\right]$ in $\epsilon_{i,j}$ will be the same if $E\left[\tilde{x}_{i,0}\right] = E\left[\tilde{x}_{i,1}\right] \ldots = E\left[\tilde{x}_{i,J}\right]$, or if the mean of the product characteristics is the same across products $0, 1, \ldots, J$. Differences in utility, $u_{i,j} - u_{i,0}$, govern discrete choices, and adding the same constant $\left(\beta_i - E\left[\beta_i\right]\right) E\left[\tilde{x}_{i,j}\right]$ to the utility of each choice will not affect $\Pr\left(j \mid x_i;\, f, \alpha\right)$.

Therefore, the use of centering in our specification is based on two alternative assumptions, 1) that $E\left[\tilde{x}_{i,0}\right] = E\left[\tilde{x}_{i,1}\right] \ldots = E\left[\tilde{x}_{i,J}\right]$, or 2) that the utility is primitively specified as

$$u_{i,j} = \alpha_j + \beta_i x_{i,j} + \epsilon_{i,j},$$

with $E\left[x_{i,j}\right] = 0$ and $\epsilon_{i,j}$ having the type I extreme value distribution. The second assumption is just as arbitrary as assuming that the utility functions without centered characteristics have type I extreme value distributed errors. There is no criterion in economic theory to prefer centered or non-centered characteristics. The paper proceeds with using centered product characteristics.

B Proofs of the Theorems

B.1 Proof of Theorem 3.1

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 $\nabla^{\upsilon} h(w) = \frac{\partial^{\upsilon} h(w)}{\underbrace{\partial w' \otimes \partial w' \otimes \ldots \otimes \partial w'}_{\upsilon \text{ Kronecker product of } \partial w'}}$.

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 3.1 (ii), for any $\upsilon = 1, 2, \ldots$ and the chain rule repeatedly applied to the linear index $x'\beta$,

$$\begin{aligned} \nabla^{\upsilon} P(x)\big|_{x=0} &= \int g^{(v)}\left(x'\beta\right)\Big|_{x=0} \left\{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\right\} f(\beta) d\beta \qquad (6) \\ &= g^{(v)}(0) \int \left\{\beta' \otimes \beta' \otimes \cdots \otimes \beta'\right\} f(\beta) d\beta. \end{aligned}$$

For each υ there are K^{υ} equations. Recall g is a known function. Therefore, as long as $g^{(v)}(0)$ is nonzero and finite for all $\upsilon = 1, 2, \ldots.,$ we obtain the υ-th moments of β for all $\upsilon \geq 1$. Now by Assumption 3.1, f satisfies the Carleman condition. Therefore, f is identified since a probability measure satisfying the Carleman condition is uniquely determined by its moments.

B.2 Proof of Theorem 4.1

Identification arises from identifying all moments, as in Theorem 3.1. We wish to show that the set $\mathcal{A}$ as defined in (4) has measure 1 in $\mathbb{R}^J$.

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

$$\begin{aligned} D_c g_j(c; \alpha) &= \frac{1}{(A_j + e^c)^2} A_j e^c, \; D_c^2 g_j(c; \alpha) = \frac{1}{(A_j + e^c)^3} \left(A_j^2 e^c - A_j e^{2c} \right) \\ D_c^3 g_j(c; \alpha) &= \frac{1}{(A_j + e^c)^4} \left(A_j^3 e^c - 4 A_j^2 e^{2c} + A_j e^{3c} \right), \ldots \end{aligned}$$

For $p \geq 3$, now we write the $(p-1)$-th derivative as

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

Then, we can write the p-th derivative as

$$
\begin{aligned}
D_c^p g_j(c;\,\alpha) &= \left[\frac{1}{(A_j+e^c)^{p+1}}\sum_{l=1}^{p}\gamma_{p+1-l}^{(p+1)}A_j^{p+1-l}e^{lc}\right] &(7)\\
&= D_c D_c^{p-1} g_j(c)\\
&= D_c\left[\frac{1}{(A_j+e^c)^p}\sum_{l=1}^{p-1}\gamma_{p-l}^{(p)}A_j^{p-l}e^{lc}\right]\\
&= \frac{1}{(A_j+e^c)^p}\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p-l}e^{lc}-\frac{1}{(A_j+e^c)^{p+1}}p\sum_{l=1}^{p-1}\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\\
&= \frac{1}{(A_j+e^c)^{p+1}}\left((A_j+e^c)\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p-l}e^{lc}-\sum_{l=1}^{p-1}p\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\right)\\
&= \frac{1}{(A_j+e^c)^{p+1}}\left(\begin{array}{c}\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p+1-l}e^{la}+\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\\ -\sum_{l=1}^{p-1}p\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\end{array}\right)\\
&= \frac{1}{(A_j+e^c)^{p+1}}\left(\begin{array}{c}\gamma_{p-1}^{(p)}A_j^pe^c+\sum_{l'=2}^{p-1}l'\gamma_{p-l'}^{(p)}A_j^{p+1-l'}e^{l'c}\\ +\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}-\sum_{l=1}^{p-1}p\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\end{array}\right) &(8)\\
&= \frac{1}{(A_j+e^c)^{p+1}}\left(\begin{array}{c}\gamma_{p-1}^{(p)}A_j^pe^c+\sum_{l=1}^{p-2}(l+1)\,\gamma_{p-l-1}^{(p)}A_j^{p-l}e^{(l+1)c}\\ +\sum_{l=1}^{p-1}l\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}-\sum_{l=1}^{p-1}p\gamma_{p-l}^{(p)}A_j^{p-l}e^{(l+1)c}\end{array}\right) &(9)\\
&= \frac{1}{(A_j+e^c)^{p+1}}\left(\begin{array}{c}\gamma_{p-1}^{(p)}A_j^pe^c-\gamma_1^{(p)}A_j^1e^{pc}\\ +\sum_{l=1}^{p-2}\left\{(l+1)\,\gamma_{p-l-1}^{(p)}+l\gamma_{p-l}^{(p)}-p\gamma_{p-l}^{(p)}\right\}A_j^{p-l}e^{(l+1)c}\end{array}\right) &(10)
\end{aligned}
$$

where in (8) and (9), we take out the first element in the first sum and change the index l' to $l+1$. (10) 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 (7) and (10) and obtain

$$
\begin{aligned}
\sum_{l=1}^{p}\gamma_{p+1-l}^{(p+1)}A_j^{p+1-l}e^{lc} &= \gamma_p^{(p+1)}A_j^pe^c+\sum_{l=2}^{p-1}\gamma_{p+1-l}^{(p+1)}A_j^{p+1-l}e^{lc}+\gamma_1^{(p+1)}A_j^1e^{pc}\\
&= \gamma_p^{(p+1)}A_j^pe^c+\sum_{l=1}^{p-2}\gamma_{p-l}^{(p+1)}A_j^{p-l}e^{(l+1)c}+\gamma_1^{(p+1)}A_j^1e^{pc}\\
&= \gamma_{p-1}^{(p)}A_j^pe^c+\sum_{l=1}^{p-2}\left\{(l+1)\,\gamma_{p-l-1}^{(p)}+l\gamma_{p-l}^{(p)}-p\gamma_{p-l}^{(p)}\right\}A_j^{p-l}e^{(l+1)c}-\gamma_1^{(p)}A_j^1e^{pc}.
\end{aligned}
$$

We find

$$\gamma_p^{(p+1)} = \gamma_{p-1}^{(p)} \tag{11}$$

$$\gamma_{p-l}^{(p+1)} = (l+1)\,\gamma_{p-l-1}^{(p)} - (p-l)\,\gamma_{p-l}^{(p)} \text{ for } p \geq 3 \tag{12}$$

$$\gamma_1^{(p+1)} = -\gamma_1^{(p)}. \tag{13}$$

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

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

and when $p = 3$, we find

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

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

$$\sum_{l=1}^{p} \gamma_{p+1-l}^{(p+1)} A_j^{p-l} = 0. \tag{14}$$

We consider two cases. In the first case where $\alpha_j = 0$ for all j, i.e., all goods are symmetric, the condition of nonzero derivatives is trivially obtained using the well-known “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 (14) is equal to $\gamma_p^{(p+1)} = 1$ for all p. Also note that the constant term (the coefficient on A_j^0 in (14)), $\gamma_1^{(p+1)}$, is equal to 1 when p is odd and is equal to -1 when p is even. By the well-known “rational zero test”, this implies that the only possible positive rational number solution in (14) is $J = 1$. A positive integer greater than 1 cannot be the solution of (14) for any p. This proves the condition of nonzero derivatives for the symmetric case.

Second we consider the asymmetric case where α_j may not be 0. In this case we allow for non-integer solutions of the equation (14).[9] Nonetheless we will show that the set

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

has measure zero in $\mathbb{R}^J$.

Note that the set of values $\mathbb{A} \subset \mathbb{R}_+$ that collect values of A_j that solve the equation (14) for at least one p is countable because for any order p the equation (14) 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 $\tilde{\mathcal{A}}(\tilde{A}) = \{\alpha \mid A_j(\alpha) = \tilde{A}, \alpha \in \mathbb{R}^J\}$ has measure zero in $\mathbb{R}^J$ because $\tilde{\mathcal{A}}$ is at most a subset of the vector space of $\mathbb{R}^{J-1}$. Next note that by construction we have $\mathcal{A}^C = \cup_{\tilde{A}\in\mathbb{A}}\tilde{\mathcal{A}}(\tilde{A}) = \cup_{\tilde{A}\in\mathbb{A}}\{\alpha \mid A_j(\alpha) = \tilde{A}, \alpha \in \mathbb{R}^J\}$. Finally we conclude that the set $\mathcal{A}^{\mathcal{C}}$ has measure zero in $\mathbb{R}^J$ because a countable union of measure zero sets has measure zero noting that $\tilde{\mathcal{A}}(\tilde{A})$ has measure zero for all $\tilde{A} \in \mathbb{A}$ and the set $\mathbb{A}$ is countable. This completes the proof.

[9]Because the equation (14) holds for any j, in practice we may pick j such that $j^* = \text{argmin}\{\alpha_1, \ldots, \alpha_J\}$. Then $A_{j^*} = \exp(-\alpha_{j^*}) + \sum_{j' \neq j^*} \exp(\alpha_{j'} - \alpha_{j^*}) > J - 1$. This again rules out all rational roots of (14) with $J \geq 2$ because the only rational root of (14) is $A_{j^*} = 1$ according to the "rational zero test".

If A_{j^*} is not a solution of (14) for any order of derivative, then we satisfy the nonzero derivatives condition in Assumption 3.2. Suppose A_{j^*} is a solution to the equation (14) corresponding to the p-th order derivative. Then we cannot identify the p-th moments of $f(\beta)$ using our strategy because $D_c^p g_{j^*}(c = 0) = 0$. But we can also utilize the share equation of other good (including the outside good) to identify the p-th moments of $f(\beta)$. We can choose a j'-th good such that (i) $A_{j'} = \exp(-\alpha_{j'}) + \sum_{j \neq j'} \exp(\alpha_j - \alpha_{j'}) \neq A_{j^*}$ and (ii) $A_{j'}$ does not solve the equation (14) corresponding to the p-th order derivative. Then we can identify the p-th order moments of $f(\beta)$ from the share equation of the j'-th good. Therefore, combining these two share equations, we can identify all the moments of $f(\beta)$.

Moreover, we conjecture (we have experimented this) that if A_{j^*} solves (14) for the p-th order derivative, A_{j^*} cannot be a solution to (14) for any other order derivative than the p-th order. Therefore all the moments of $f(\beta)$ other than the p-th order moments are identified using derivatives of the share equation for the j^*-th good.

References

[1] Akhiezer, N.I. (1965), *The classical moment problem and some related questions in analysis*, Oliver & Boyd.

[2] Bajari, P., J. Fox, K. Kim, and S. Ryan (2009), "A Simple Estimator for the Distribution of Random Coefficients", NBER working paper w15210.

[3] Berry, S and Haile, P. (2008), "Nonparametric Identification of Multinomial Choice Models with Heterogeneous Consumers and Endogeneity", working paper.

[4] Boyd, JH and RE Mellman (1980), "Effect of Fuel Economy Standards on the U. S. Automotive Market: An Hedonic Demand Analysis", *Transportation Research B*, 14(5), 367–378.

[5] Briesch, Richard A., Pradeep K. Chintagunta, and Rosa L. Matzkin (2009), "Nonparametric Discrete Choice Models with Unobserved Heterogeneity", *Journal of Business and Economic Statistics.*

[6] Burda, Martin, Matthew Harding and Jerry Hausman (2008), "A Bayesian Mixed Logit-Probit for Multinomial Choice", *Journal of Econometrics*, 147(2), 232–246.

[7] Cardell, NS and FC Dunbar (1980), "Measuring the societal impacts of automobile downsizing", *Transportation Research B*, 14(5), 423–434.

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

[9] Chiappori, PA and I. Komunjer (2009), "On the Nonparametric Identification of Multiple Choice Models", Columbia University working paper.

[10] Fox, Jeremy T. and Amit Gandhi (2009), "Identifying Heterogeneity in Economic Choice Models", University of Chicago working paper.

[11] Hausman, Jerry and David A. Wise (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences", *Econometrica,* 46(2), 403–426.

[12] Heckman, JJ and RJ Willis (1977), "A beta-logistic model for the analysis of sequential labor force participation by married women", *Journal of Political Economy*, 85(1), 27–58.

[13] 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.

[14] Kim, Kyoo il, Amil Petrin and Kenneth Train (2009), "Control Function Corrections for Unobserved Factors in Differentiated Product Models", University of Minnesota working paper.

[15] Krein, M.G. and A.A. Nudel'man (1973, translation 1977), *The Markov moment problem and extremal problems: ideas and problems of P. L. Čebyšev and A. A. Markov and their further development.* Translations of mathematical monographs v. 50, American Mathematical Society, Providence.

[16] Lewbel, Arthur. (2000) "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables", *Journal of Econometrics*, 97, 1, 145-177.

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

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

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

[20] Train, K.E. (2009), "EM Algorithms for Nonparametric Estimation of Mixing Distributions", *Journal of Choice Modelling*, 1(1), 40–69.