Nonparametric Identification and Estimation of Random Coefficients in Nonlinear Economic Models

Jeremy T. Fox
University of Chicago and NBER

Amit Gandhi
University of Wisconsin-Madison*

January 2010

Abstract

We show how to nonparametrically identify and estimate the distribution of random coefficients that characterizes the heterogeneity among agents in a general class of economic choice models. We introduce an axiom that we term separability and prove that separability of a structural model ensures identification. Identification naturally gives rise to a nonparametric minimum distance estimator. We prove identification of distributions of utility functions in multinomial choice, distributions of labor supply responses to tax changes, and distributions of wage functions in the Roy selection model. We also reconsider the problem of endogeneity in economic choice models, leading to new results on the two-stage least squares model.

*Thanks to Stephane Bonhomme, Steven Durlauf, James Heckman, Salvador Navarro, Philip Reny, Susanne Schennach, Azeem Shaikh, Christopher Taber, Harald Uhlig and Edward Vytlacil for helpful comments. Also thanks to seminar participants at Boston University, Brown, the Brown / UCL Demand Conference, Caltech, CREST, Chicago, Cowles, EC2 Rome, Georgetown, Indiana, LSE, Michigan, Michigan State, Minnesota, Northwestern, Penn State, Pittsburgh, Rochester, the SED, Stanford, Toulouse, UCL, USC, Washington University in St. Louis, Wisconsin and Yale. Fox thanks the National Science Foundation, the Olin Foundation, and the Stigler Center for financial support. Thanks to Chenchuan Li for research assistance. Our email addresses our fox@uchicago.edu and agandhi@ssc.wisc.edu.

1 Introduction

Heterogeneity among decision makers, be they firms or consumers, is a critical feature of economic life that is important for the study of many policy problems. A classic example is models of demand for differentiated products. We observe many competing products, each with different characteristics, being offered for sale. The presence of these differentiated products suggest that consumers have heterogeneous preferences for product characteristics. In order to model the demand for these products, we need to estimate the distribution of consumers' heterogeneous preferences.

Likewise, in the classic Roy model, a worker chooses between different sectors of employment based on the wage offered to that worker in each sector. As we observe workers picking different sectors, it is likely that workers have heterogeneous wage outcomes in the various sectors. We observe in the data only the wage of the sector the worker picked. To model substitution between sectors as circumstances change, we need to know the distribution of the heterogeneous wage functions in each of the sectors. The Roy model is mathematically similar to the competing risks model, where an agent dies from the deadliest of several diseases he might have. It is likely the response to diseases may be heterogeneous across agents.

In both the differentiated products demand model and the Roy model, the key unknown object to researchers is a distribution of heterogeneity. In demand, it is the distribution of preferences for product characteristics. In the Roy model, it is the distribution of wage outcome functions for each of the sectors. This paper presents tools for identifying and estimating the distributions of heterogeneity in a class of economic models that includes the multinomial choice demand model and the Roy selection model. We focus on the common data scheme in applied microeconomics where the researcher has access to cross-sectional rather than panel data. For example, in a demand environment, we observe different agents making choices at different budget sets. In the Roy model, we observe different workers making employment choices in different labor markets. We will characterize each agent in the population by a behavioral parameter $\theta \in \Theta$, which we will refer to as a "random coefficient" as θ is heterogeneous in the population with distribution G. We seek the nonparametric identification and estimation of G.

The nonparametric identification and estimation of distributions of random coefficients is well-understood in the case of the linear regression model with random coefficients. Let $y = a + x'b$, where x is a real vector with continuous support and y is a real vector of outcome variables. In the random coefficient model, a and b are heterogeneous vectors of intercept and slope parameters, respectively. One example is a Cobb-Douglas production function in logs, where y is log output, x is a vector of log inputs, a is total factor productivity, and b is the vector of input elasticities. Compared to a model without random coefficients, the random coefficient regression model allows the effect of changing inputs x to vary across firms: b is a heterogeneous parameter. Some firms in the same industry may have labor-intensive technologies and others may have capital-intensive technologies. In the random coefficient production function model, the object of interest is the distribution of random coefficients $G(a,b)$, or the joint distribution of total factor productivities and input elasticities. Knowledge of $G(a,b)$ tells us the distribution of production functions in an industry, which is an argument to answering many policy questions, for example the effects of taxing some particular input. In a seminal paper, Beran and Hall (1992) first demonstrated the non-parametric identification of the distribution of the slope and intercept parameters in the linear model. In follow-up work, Beran and Millar (1994) and Hoderlein, Klemelä and Mammen (forthcoming) investigate the nonparametric estimation of $G(a,b)$ in the linear model.

The only attempt at extending identification to a nonlinear setting is Liu (1996). He studies the model $y = g(x,\theta)$, where $g(\cdot,\cdot)$ is a known continuous function, θ is a finite vector of real parameters, and, as above, y is a real vector of outcome variables and x is a real vector with continuous support. For example, $g(\cdot,\cdot)$ might represent some nonlinear production function, such as the constant elasticity of substitution (CES). The object of interest is $G(\theta)$, the distribution of the heterogeneous parameters or random coefficients. Although he writes that "general results for the nonlinear model seem impossible", he does show that some traction on the problem is achieved under two main regularity conditions, which we will also adopt to some extent, but will also generalize. The key regularity conditions in Liu are that the underlying distribution G over Θ in the population is a multinomial distribution, and further that the component functions g_i that constitute $g = (g_1, \ldots, g_m)$ are each real analytic in x. Under these assumptions, he shows that the distribution G can be identified from the joint distribution of (y,x).

Despite generalizing the linear in random coefficients model to a nonlinear setting, Liu's results are still limiting from the point of view of economic application. An important assumption in Liu is that the underlying relationship between y and x for any $\theta \in \Theta$ is continuous. Unfortunately for applied work, requiring the relationship between y and x to be continuous is a major constraint. Discrete choice and the Roy selection models are altogether ruled out, as each of these economic mechanisms induces a discontinuity between the response and the covariates. However it is for these very class of problems that identification of unobserved heterogeneity has been the most critical from a policy perspective, as we argued before. Additionally, Liu requires that the regressors x are distributed independently of the random coefficients θ, thus ruling out any potential endogeneity of the regressors. Endogeneity is of course a fundamental concern in economic applications.

We show that distribution of the random coefficients, i.e., the distribution G, can be nonparametrically identified in a much wider class of economic models. We first assume independence between the regressors x and the random coefficients θ, but allow for a discontinuous economic relationship between the outcome y and the regressors x. In particular we consider a general class of models in the form of $y = f(x,\theta) = \Upsilon(g(x,\theta), x)$, where the vector-valued function g is a "regular" continuous functions (we will use a weaker form of the real analytic assumption to define "regular"), and the transformation Υ is the source of the discontinuity between y and x. The Roy selection model, censored regression models, and multinomial choice demand models all have such a representation. In the event that $f(x,\theta) = g(x,\theta)$, then we are back to the setup considered in Liu. To our knowledge, we are the first to prove that the joint distribution of random, choice-specific utility functions is identified in a multinomial choice model. Likewise, we prove that the joint distribution of random, sector-specific return functions is identified in the Roy selection model.

While we initially proceed under the classical assumption that x and θ are independent in the population, we can allow for endogeneity among a subset of regressors that is modeled as a triangular system of equations with instruments. Our results provide important economic generalizations of the classic two-stage least squares (2SLS) model. For example, we allow the response of an endogenous regressor to an instrument to be heterogeneous across the population. Also, there are no monotonicity assumptions on the response to the instrument. We also are one of the first papers to address endogeneity in multinomial choice models in a nonparametric setting. Demand estimation is a widely-used tool in industrial organization, marketing and other fields. We show how to use a triangular system and instruments to address price endogeneity.

While the results above are attained assuming a sufficient regularity in the underlying functional relationships $g(x,\theta)$ that exhibit heterogeneity in the population (that we capture through a generalization of real analytic functions), for certain policy questions, this degree of regularity is not required. For example, we consider an extension of Liu to the case of continuous response. Fixing a point of interest $x^\star$, we are able to identify the distribution of marginal effects at $x^\star$, or the distribution of $\left.\frac{\partial g(x,\theta)}{\partial x}\right|_{x=x^\star}$ when θ is a random variable. A marginal effect is a causal policy parameter, for example the effect of taxes on hours of worked at a tax level $x^\star$. We identify not just the mean labor supply response to the tax change, but the distribution of responses. Unlike Liu and our results on multinomial choice, to identify the distribution of marginal effects we do not rely on assuming that the $g(x,\theta)$ are real analytic functions.

Having shown identification for a general class of nonlinear models, we can adopt the Beran and Millar (1994) nonparametric minimum distance estimator and show that the estimator continues to be consistent in the nonlinear random coefficients setting. In fact the multinomial assumption on the distribution of heterogeneity G has an affinity with the minimum distance estimator, as the estimator maximizes a minimum distance criterion over a set of multinomial distributions. As noted by Foster and Hahn (2000) in their application of Beran and Millar (1994), each support point of the multinomial distribution has the natural interpretation as a "type" in the population, and the multinomial assumption is thus that there are a finite number of types in the population whose number, location, and masses need to be identified from the data.

2 The Identification Problem

We consider a general class of economic models where each model $\mathcal{M}$ can be described by a tuple $\mathcal{M} = (\Theta, \mathcal{X}, \mathcal{Y}, f)$. The set Θ denotes a functional space representing the feasible set of types of agents admitted by the model. The set $\mathcal{X}$ denotes the set of economic environments in the support of the data generating process. The set $\mathcal{Y}$ is the (measurable) outcome space. The function $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ maps an agent's type $\theta \in \Theta$ and economic environment $x \in \mathcal{X}$ to an outcome $y = f(x, \theta) \in \mathcal{Y}$. The joint distribution of outcomes and environments (y, x) is identified from the i.i.d. data. What remains to be identified is the distribution of types $G \in \mathcal{G}$ in the population, where $\mathcal{G}$ is a set of probability measures over Θ.[1]

Let $A \subseteq \mathcal{Y}$ be a measurable subset of the outcome space. Assuming stochastic independence between the structural error θ and the covariates x, if $G^0 \in \mathcal{G}$ is the true distribution of types in the population, we have that

$$\Pr_{G^0}(A \mid x) = G^0(\{\theta \in \Theta \mid f(x, \theta) \in A\}) = \int 1[f(x, \theta) \in A]\, dG^0(\theta). \tag{1}$$

Thus the distribution G^0 is identified up to the measure it assigns to sets of the form $I_{A,x} = \{\theta \in \Theta \mid f(x, \theta) \in A\}$, which are indexed by a point x and a set $A \subseteq \mathcal{Y}$. The problem is whether the class of such sets $I_{A,x}$ is rich enough to point identify G^0 within a class of distributions $\mathcal{G}$.

To state this problem precisely, let $\Pr(\cdot \mid x)$ be a probability measure over $\mathcal{Y}$ for a given value $x \in \mathcal{X}$ of the environment. Let $P = \{\Pr(\cdot \mid x) \mid x \in \mathcal{X}\}$ denote a collection of such probability measures over all possible economic environments and let $\mathcal{P}$ denote the set of all such collections P. Then we can view (1) as a mapping $L : \mathcal{G} \to \mathcal{P}$. We will say the model $\mathcal{M}$ is *identified* relative to $\mathcal{G}$ if L is a one-to-one map. That is, for any $G, G' \in \mathcal{G}$ and $G \neq G'$, there exists an experiment in the data (A, x) where $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$ such that $\Pr_G(A \mid x) \neq \Pr_{G'}(A \mid x)$, where $\Pr_G(\cdot \mid x)$ and $\Pr_{G'}(\cdot \mid x)$ are the images of G and G' respectively under L.

The question behind the identification problem is whether the same economic population G facing exogenously varying economic environments $x \in \mathcal{X}$ will have revealed preferences, in the form of the reduced form relationship in the data $\Pr(A \mid x)$, that are informative enough to identify G. Mathematically, the identification problem can be understood as an existence problem. Identification requires showing that, for any two potential distribution of types, there always exists an experiment in the data (A, x) that can empirically distinguish between these distributions. In the next section, we show that there exists such an experiment (A, x) if the economic model $\mathcal{M}$ satisfies a separability condition.

[1] In parametric models, the type space Θ is a finite dimensional space. In the application of our main result, we will treat the type space Θ as an infinite dimensional functional space that nests all "regular" finite dimensional functional forms (where "regular" functional forms will be a slight generalization of analytic functions and thus encompass a wide range of economic relationships). For the development of the general theory, however, we make no explicit use of any structure on Θ and thus treat it as an arbitrary type space.

We focus on nonparametric identification, which in our context means that we do not put any parametric structures on either the type space Θ or the set of distributions $\mathcal{G}$. A lack of nonparametric identification calls into question any parametric estimator of the model: apparently the parametric estimator is only consistent because of parametric functional form restrictions either on the types θ or the distribution G. Our main restriction is that we take $\mathcal{G}$ to be the class of all finite distributions over Θ. Thus the restriction being placed on the distribution of types $G \in \mathcal{G}$ is that the set of types having positive support in the population is at most finite. However, the number of support points, the location of the support points and their masses are a priori unknown and need to be identified from the data. Thus $\mathcal{G}$ constitutes an infinite dimensional space of distributions. The class $\mathcal{G}$ can be defined without requiring any a priori structure on Θ, thus allowing us to be nonparametric about the type space Θ.[2] As will be demonstrated later, the ability to be nonparametric about Θ allows for the general applicability of our results to specific economic contexts. Variations on the discrete mixtures assumption (finite with known number of types, finite with unknown number of types, countable with known support, countable with unknown support) are common in the statistics and econometrics literatures (Teicher, 1963; Yakowitz and Spragins, 1968; Liu, 1994; Kasahara and Shimotsu, 2008; Allman et al., 2009; Hu and Shum, 2009).[3]

3 Separability

Recall the basic question is whether the class of sets of the form $I_{A,x} = \{\theta \in \Theta \mid f(x,\theta) \in A\}$ generated by the model $\mathcal{M}$ is rich enough to identify G^0 within the class of finite distributions $\mathcal{G}$. We now show that an affirmative answer to this question holds under a condition on $\mathcal{M}$ that we term separability. Separability is a strengthening of what is clearly a necessary condition for identification: for any two types θ and θ', there exists an $A \subset \mathcal{Y}$ and $x \in \mathcal{X}$ such that $f(x,\theta) \in A$ and $f(x,\theta') \notin A$, i.e., θ and θ' can be separated by (A, x). In order to state separability formally, we first define I-sets, which are objects that play a critical role in the remainder of the paper.

Definition 3.1. For any set of types $T \subset \Theta$, and for any $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$, the I-set $I^T_{A,x}$ is defined as

$$I^T_{A,x} \equiv \{\theta \in T \mid f(x,\theta) \in A\}.$$

An I-set is the set of types within an arbitrary subset of types $T \subset \Theta$ whose response is in the set A

[2] This contrasts with the class of distributions that admit density functions, which is non-nested with the class of discrete distributions, and would have to be defined contingent on the measurability properties of the underlying space Θ. This is difficult to do with general infinite-dimensional spaces.

[3] The space of countable (indeed finite) distributions is dense in the space of all probability measures over Θ so long as Θ is a metrizable topological space (Aliprantis and Border, 2006, Theorem 15.10).

at the covariates x. The key feature of I-sets is that they are strictly a property of the underlying economic choice model $\mathcal{M}$ (and are independent of the particular distribution of heterogeneity G). Our main result shows that if I-sets exhibit enough variation, then identification is achieved.

Definition 3.2. The model $\mathcal{M}$ is **separable** if, for any finite set of types $T \subset \Theta$, there exists a pair (A, x) such that corresponding I-set $I^T_{A,x}$ is a singleton.

In the definition, $T \subset \Theta$ can be any arbitrary, finite subset. The full set of feasible types Θ within the model is typically an uncountably infinite set that is quite distinct from the finite subsets T considered in the definition. We now state and prove our main result.

Theorem 3.3. *If the model $\mathcal{M}$ is separable, then the model is identifiable with respect to the class of finite distributions.*

Proof. Recall that identification requires showing that the mapping $L : \mathcal{G} \to \mathcal{P}$ defined by (1) is one to one. Thus for $G^0, G^1 \in \mathcal{G}$ with $G^0 \neq G^1$, we must have that $\Pr_{G^0}(A \mid x) \neq \Pr_{G^1}(A \mid x)$ for some $A \subseteq \mathcal{Y}, x \in \mathcal{X}$. In particular, for any point $P \in L(\mathcal{G})$, we show that $L\left(G^0\right) = L\left(G^1\right) = P$ implies $G^0 = G^1$.

Observe that we can represent any $G \in \mathcal{G}$ by a pair (T, p), where $T = \{\theta_1, \ldots, \theta_n\} \subset \Theta$ is a finite set of types and the probability vector $p = \{p_\theta\}_{\theta \in T}$ comprises non-negative masses that sum to one over T. Given the representation (T, p) for $G \in \mathcal{G}$, we can express (1) as

$$\Pr_G(A \mid x) = \sum_{\theta \in I^T_{A,x}} p_\theta. \tag{2}$$

If G^0 is represented by (T^0, p^0) and G^1 is represented by (T^1, p^1), then we can redefine p^0 and p^1 so that G^0 and G^1 are represented by (T, p^0) and (T, p^1) respectively, where $T = T^0 \cup T^1$ (for example, if $\theta \in T - T^0$, then set $p^0_\theta = 0$). T is still finite. Moreover if we define the vector $\{\pi_\theta\}_{\theta \in T}$ such that $\forall \theta \in T, \pi_\theta = p^0_\theta - p^1_\theta$, then $G^0 = G^1$ if and only if $\pi_\theta = 0$ for all $\theta \in T$.

Our goal is to show that $L\left(G^0\right) = L\left(G^1\right)$ implies $G^0 = G^1$. Observe that $L\left(G^0\right) = L\left(G^1\right)$ implies that for all $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$, $\Pr_{G^0}(A \mid x) = \Pr_{G^1}(A \mid x) = \Pr(A \mid x)$, which by (2) implies that

$$\sum_{\theta \in I^T_{A,x}} \pi_\theta = 0, \tag{3}$$

for all I-sets $I^T_{A,x}$. We now show that (3) implies $\pi_\theta = 0$ for all $\theta \in T$. Assume to the contrary that $T_2 = \{\theta \in T \mid \pi_\theta \neq 0\}$ is non-empty. By separability, we can produce a singleton $I^{T_2}_{A,x} = \{\theta^\star\}$. Furthermore, we can re-write (3) as

$$\sum_{\theta \in I^T_{A,x}} \pi_\theta = \sum_{\theta \in I^{T_2}_{A,x}} \pi_\theta + \sum_{\theta \in I^{T-T_2}_{A,x}} \pi_\theta = \sum_{\theta \in I^{T_2}_{A,x}} \pi_\theta = \pi_{\theta^\star} \neq 0,$$

which contradicts (3). Hence it must be that T_2 is empty, and thus $\pi_\theta = 0$ for all $\theta \in T$. $\square$

The above theorem is properly viewed as an existence theorem, and asserts that under separability of the model, an identifying experiment (A, x) must always exist.[4]

The proof of Theorem 3.3 can be adapted almost without change for the case where separability applies to countable sets $T \subset \Theta$. We do not pursue that here because considering arbitrary countable sets T makes verifying separability more difficult in common economic models used in structural work. A finite distribution is indexed by the number of support points, the identity of each support point, and the mass of each support point. The theorem states that the researcher can identify the number, identity and the mass of the support points. As the number of support points of an element of the class of finite distributions can be arbitrarily large, it is not possible to reject the finite support assumption with a finite dataset.[5]

4 Simple Examples

This section presents two examples, one of a model that is clearly identified and one of a model that is clearly unidentified. We show that the former model satisfies separability and the latter model does not.

Let the model by $y = \theta$, with $\theta \in \Theta \subseteq \mathbb{R}$. This model is clearly identified because the distribution of y is the distribution of θ. This model also satisfies the sufficient condition of separability. Let $I_y^T = \{\theta \in T \mid \theta = y\}$. Then choosing any $y^\star = \theta^\star$ for any $\theta^\star \in T$ gives I_y^T equal to just the singleton $\{\theta^\star\}$.

Now consider the model $y = \theta_a + \theta_b$, with $\theta = (\theta_a, \theta_b) \in \mathbb{R}^2$. This model is clearly not identified and thus it is instructive to consider where separability fails. Consider the I-set $I_y^T = \{\theta \in T \mid \theta_a + \theta_b = y\}$. For a counterexample to the sufficient condition of separability, let $T = \left\{\theta^1, \theta^2\right\}$ such that $\theta_a^1 + \theta_b^1 = \theta_a^2 + \theta_b^2$ but the types are distinct, so $\theta_a^1 \neq \theta_a^2$. For $y^\star = \theta_a^1 + \theta_b^1$, $I_{y^\star}^T = T$ and for any $y \neq y^\star$, $I_y^T = \emptyset$. Thus, separability is not satisfied.

[4] The identification is non-constructive in the sense that it does not attempt to recover the underlying distribution over types (T, p) from the distribution of the data $P = \{\Pr(\cdot \mid x) \mid x \in \mathcal{X}\}$. That is, we do not consider a structure (T, p) to be the value of a functional $\mathcal{H}(P)$ of the data P (which is a typical approach used in the nonparametric identification literature because it ties identification to an analog estimator, see, e.g., Chesher (2003)). Rather the theorem shows the weaker result that the mapping $L : \mathcal{G} \to \mathcal{P}$ is injective. But this is the defining property of nonparametric identification: different structures have different observable implications.

[5] The class of finite distributions $\tilde{\mathcal{G}}$ over any infinite-dimensional set Θ is an infinite-dimensional space. Assume to the contrary that the space $\tilde{\mathcal{G}}$ was instead k-dimensional for a finite integer k. Then any $k+1$ elements of $\tilde{\mathcal{G}}$ would be linearly dependent. Let δ_θ denote the Dirac delta probability measure that assigns mass 1 to $\theta \in \Theta$. Because Θ is an infinite set, we can always find $k+1$ elements of Θ, say $\{\theta_1, \ldots, \theta_{k+1}\}$, and as a result we can always find $k+1$ elements of $\tilde{\mathcal{G}}$, namely $\{\delta_{\theta_1}, \ldots, \delta_{\theta_{k+1}}\}$. However $\{\delta_{\theta_1}, \ldots, \delta_{\theta_{k+1}}\}$ can never be a linearly dependent set. Thus $\tilde{\mathcal{G}}$ must be infinite dimensional.

5 Strong Identification and Minimum Distance Estimation

We now show that our definition of identification along with an additional regularity condition on the model implies the traditional definition of identification as employed in the econometrics literature (see e.g., Manski (1994)). Under this usual definition, we say a model is identified if the joint distribution of the observable random variables uniquely pins down the underlying structural object. In the present case, the joint distribution of (x, y) is observed in the data. Furthermore, for any $G \in \mathcal{G}$, the model predicts a joint probability measure of the random vector (x, y). In particular, for any measurable sets $B \subseteq \mathcal{X}$ and $A \subseteq \mathcal{Y}$, the model's prediction of the probability of the event $B \times A$ is

$$\mathcal{L}\left(G, F_X\right)(B, A) = \int_{x \in B} \int_{\theta \in \Theta} 1\left[f(\theta, x) \in A\right] dG(\theta)\, dF_X(x) = \int_{x \in B} \Pr_G(A \mid x)\, dF_X(x), \qquad (4)$$

where F_X is the marginal distribution of x. We will say the model is strongly identified if $\mathcal{L}(\cdot, F_X)$ as a function over the space of distributions $\mathcal{G}$ is one to one.

We now show the additional structure under which identification (as defined in the previous sections) implies this stronger form of identification. For any $G, G' \in \mathcal{G}$, simple identification tells us that there exists an identifying experiment $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$ such that

$$\Pr_G(A \mid x) \neq \Pr_{G'}(A \mid x),$$

or in the language of I-sets,

$$\sum_{i \in I^T_{A,x}} q_i \neq 0,$$

for $q_i = G(\theta_i) - G'(\theta_i)$ and $T = \{\theta_1, \ldots, \theta_n\}$ being the union of the supports of G and G'. If, in addition to the existence of the identifying experiment (A, x), we also have a neighborhood B of x such that $I^T_{A,z} = I^T_{A,x}$ for all $z \in B$, then it is straightforward to show using (4) that

$$\mathcal{L}\left(G, F_X\right)(B, A) \neq \mathcal{L}\left(G', F_X\right)(B, A),$$

and thus strong identification holds. We can thus formally state:

Lemma 5.1. *Identification of the model $\mathcal{M}$ implies strong identification if for any finite set of types $T \subset \Theta$ and identifying experiment (A, x), there exists a neighborhood B of x such that $I^T_{A,z} = I^T_{A,x}$ for all $z \in B$.*

For all the models we consider, strong identification follows from simple identification under the regularity conditions that we assume. For this reason, we will focus on the main substantive aspect of both types of identification, which is to show the existence of an identifying experiment

(A, x), which we pursue via separability.

5.1 The Minimum Distance Estimator

Once strong identification has been established, it is straightforward to adapt the nonparametric minimum distance estimator of Beran and Millar (1994) and Beran (1995), which was established for the linear-in-random-coefficients model, to our general setting. The minimum distance estimator minimizes the distance between the empirical distribution of x and y and the model's prediction of the joint distribution of x and y. Let Q_0 be the true distribution of x, y, $\hat{Q}_n$ the empirical distribution with n observations of (x_i, y_i), and $\mathcal{L}(G, F_X)$ the model's prediction of the joint distribution of x, y given a distribution of unobserved types G and a distribution of independent variables F_X (as expressed in (4)).

Let $d(,)$ be a distance function that metrizes weak convergence. For example, d could be the Lévy-Prokhorov metric on the space of multivariate distributions or a metric derived from the norm over characteristic functions. Let $\mathcal{G}^m$ be a class of finite distributions over Θ with m support points. The minimum distance estimator $\hat{G}_n$ is

$$\hat{G}_n \in \arg \inf_{G \in \mathcal{G}^{m(n)}} d\left[\hat{Q}_n, \mathcal{L}\left(G, \hat{F}_X^n\right)\right],$$

where $m(n)$ is any increasing function of n with $m(n) \to \infty$ as $n \to \infty$. Assuming that the operator $\mathcal{L}$ satisfies strong identification and is continuous, then we can mimic the proof in Beran (1995) to show that $\hat{G}_n$ consistently approaches the true distribution $G^0 \in \mathcal{G}$. The key feature of the minimum distance estimator is that consistency fundamentally relies on identification of the model. Linearity was only needed by these previous authors to show identification. Given our identification results, the consistency proof in Beran applies to our models, so we do not repeat the argument.

6 The No Ties Property on Function Spaces

Verifying that a choice model $\mathcal{M}$ satisfies separability in many of the structural models that follow will be related to the underlying functional space Θ or a part of that space satisfying a "no ties" property that we formalize in this section. The discussion of marginal effects in Section 10 demonstrates that separability can be used for models whose constituent functions do not satisfy the "no ties" property. However, the "no ties" property will be used in most of our models.

There are two versions of the no ties property, both a strong and a weak version, and both properties are satisfied by functional spaces that are quite commonly used in economic models. Indeed, both classes are generalizations of the class of real analytic functions. To establish some notation, for a given non-empty rectangle $\mathcal{X} \subseteq \mathbb{R}^k$, let $\mathcal{C}_{\mathcal{X}}^{k,m}$ denote the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}^m$. We will also use the assumption of continuity throughout the paper.

Definition 6.1. A set of functions $S_{\mathcal{X}}^{k,m} \subseteq \mathcal{C}_{\mathcal{X}}^{k,m}$ satisfies the strong no ties property (SNTP) if for any finite subset of functions $\{g_1, \ldots, g_n\} \subset \mathcal{S}_{\mathcal{X}}^{k,m}$ and any open $U \subseteq \mathcal{X}$, there exists a point $x \in U$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$.

The SNTP is in a specific sense a "generic" property of $\mathcal{C}_{\mathcal{X}}^{k,m}$. To see this, let $\mathcal{P}_{\mathcal{X}}^{k,m} \subset \mathcal{C}_{\mathcal{X}}^{k,m}$ denote the set of vector-valued polynomial functions over $\mathcal{X}$, i.e., $g = (g_1, \ldots, g_m) \in \mathcal{P}_{\mathcal{X}}^{k,m}$ if and only if $g_i : \mathcal{X} \to \mathbb{R}$ is a polynomial function over $\mathcal{X}$ for each $i = 1, \ldots, m$. Notice that $\mathcal{P}_{\mathcal{X}}^{k,m}$ is an infinite dimensional functional space, and it satisfies the SNTP. If $\mathcal{X}$ is closed and bounded, then by the Stone-Weierstrass theorem $\mathcal{P}_{\mathcal{X}}^{k,m}$ is dense in $\mathcal{C}_{\mathcal{X}}^{k,m}$ in the supremum norm. More generally, the set of vector-valued real analytic functions (which contains $\mathcal{P}_{\mathcal{X}}^{k,m}$) satisfies the SNTP. See Appendix A for a proof.[6]

A more general condition than the SNTP is the weak no ties property (WNTP), which relaxes the need to break ties in any open set $U \subseteq \mathcal{X}$. The WNTP holds for the space of real analytic functions.

Definition 6.2. A subset $\mathcal{W}_{\mathcal{X}}^{k,m} \subseteq \mathcal{C}_{\mathcal{X}}^{k,m}$ satisfies the weak no ties property (WNTP) if for any finite subset $\{g_1, \ldots, g_n\} \subset \mathcal{W}_{\mathcal{X}}^{k,m}$ there exists $x \in \mathcal{X}$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$.

In many but not all of the models we consider, we show separability and hence identification of choice models $\mathcal{M}$ with respect to finite distributions G by exploiting the SNTP or the WNTP on the underlying functional space of types Θ, or on portions of that space.

As the maximal set of functions that satisfy SNTP and WNTP have not been determined, invoking either property is, under current knowledge, a restriction to the class of real analytic functions.[7] One practical consequence of our identification results in the remaining sections is that they immediately apply to identification of a nonparametric distribution in parametric models specified up to a finite number of heterogeneous parameters, if that parametric model uses a functional form choice that lies in the class of real analytic functions. Commonly used production and demand functions, such as the translog, are real analytic. The class of real analytic functions excludes many functions of traditional interest, such as piecewise linear functions and splines.

[6] Real analytic functions are defined formally in the appendix, but roughly speaking, they are functions that can be locally parameterized by a countable parameter vector. The space of real analytic functions nests all polynomials of any finite order. Examples of real analytic functions include the simple functions such as exp, sin, and log, as well as algebraic combinations and compositions of these functions.

[7] Real analytic functions are differentiable and thus we restrict attention to functions of variables with continuous support. If discrete characteristics d exist, we can condition on them. In other words, we can identify a distribution $G(g \mid d)$ over functions $g(x \mid d)$, for the Liu (1996) example, for each observable value of d.

7 Multinomial Choice

Multinomial choice is a key model used in empirical industrial organization to model consumer demand. Demand functions are useful for measuring market power and predicting the welfare gain from new goods. This section shows how discrete choice models of demand are nonparametrically identified within our framework. As discrete choice introduces a discontinuity between an agent's response and the covariates, it falls outside of the scope of the nonparametric random coefficients literature to date. We will show identification using a common condition in the literature, a large support condition on one regressor.[8]

Consider an agent θ making a discrete choice from among J products and one outside good. Let $\mathcal{Y} = \{0, 1, \ldots, J\}$, where 0 is the outside good. Each product $j \in \mathcal{Y} - \{0\}$ is characterized by a scalar characteristics $w_j \in \mathbb{R}$. We let $v \in \mathbb{R}^K$ denote the observed characteristics of the consumer and the menu of product characteristics (the J products) excluding the scalar characteristics, $w = (w_1, \ldots, w_J)$. We let $x = (v, w) \in \mathbb{R}^{K+J}$ denote the entire menu of consumer and product characteristics including the scalar characteristics. We will follow the usual convention that the permissible range of variation in each w_j for $j \in J$ is independent of the product characteristics v.

Assumption 7.1. *Let $V \subset \mathbb{R}^K$, the support of v, be a non-empty rectangle. Let $x = (v, w) \in \mathcal{X} = V \times W_1 \times \cdots \times W_J$ where $W_j = \mathbb{R}$ for each $j \in J$.*

A type $\theta = u = \left(u^1, \ldots, u^J\right)$ is a vector of functions of the product characteristics $v \in V$. That is, a type is a function $u : V \to \mathbb{R}^J$. Utility functions are heterogeneous across the units of observation. The goal is to identify their distribution.

Assumption 7.2. *The function u is statistically independent of the observable choice set $x = (v, w)$.*

We discuss endogeneity in Section 9.3. To show separability, we will need a monotonicity assumption for the special regressor w_j.

Assumption 7.3. *The utility of a type u purchasing product j is $u^j(v) + w_j$.*

The quasi-linearity of the utility for choice j in w_j ensures that at any v there will be a set of w_j's where a given type u will switch to a different choice. Otherwise, it could be that a type will never to switch choices at any observable set of product characteristics. We also introduce an outside good that we label good $j = 0$ whose utility is normalized to 0 for each agent. An agent's response at $x = (v, w)$ is given by the discrete choice that maximizes utility, or

$$f(x, \theta) = \arg\max_{j \in \mathcal{Y}} \left\{ u^j(v) + w_j \right\},$$

[8] This large support condition is not related to the condition of "identification at infinity" implicitly used in Lewbel (2000), as pointed out by Magnac and Maurin (2007). We use large support on w_j to induce agents with high preferences for a particular good (at a particular v) to switch choices.

where $u^0(v) + w_0 \equiv 0$. We restrict attention to utility functions that satisfy the weak no-ties property.

Assumption 7.4. *The type space Θ of feasible utility functions satisfies the WNTP.*

Example 7.5. A special case of the model is when $u^j(v) + w_j = v_j'\beta + w_j + \epsilon_j$, where v_j is a vector of product characteristics for choice j, ϵ_j is a scalar, choice-specific unobservable, and β is a vector of marginal utilities that are heterogeneous across consumers. We let $v = (v_1, \ldots, v_J)$. The scale normalization is on w_j instead of ϵ_j, as in the logit model of McFadden (1973). The functions $u^j(v) = v_j'\beta + \epsilon_j$ are real analytic in v, and hence the class Θ satisfies the WNTP. Our theorem below, for this example, will show the nonparametric identification of the joint distribution $G(\beta, \epsilon_1, \ldots, \epsilon_J)$ under the assumption that G has a finite number of support points.

In the general model, letting the utility to product j also depend on the characteristics of products $k \neq j$ can capture the idea of context or "menu" effects in consumer choice. Even if such effects are not economically desirable, there is no cost to us in mathematical generality and thus we let the whole menu v enter as an argument to each u^j. The choice-specific scalar w_j, however, enters preferences in an additively separable way (and hence preferences are quasilinear in this scalar characteristic). One example is that w_j could be the price of good j, in which case $u^j(v)$ is type u's reservation price for product j, and preferences are better expressed as $u^j(v) - w_j$. However, w_j could be some non-price product characteristic or, with individual data, an interaction of a consumer and product characteristic, like the geographic distance between a consumer and a store.

Implicit in the quasilinear representation of preferences $u^j(v) + w_j$ is the scale normalization that each type's coefficient on w_j is constrained to be 1. The normalization of the coefficient on w_j to be ± 1 is innocuous; choice rankings are preserved by dividing any type's utilities $u^j(v) + w_j$ by a positive constant. Thus if w admitted a type-specific coefficient $\alpha > 0$, then the type (u, α) would have the exact same preferences as the type $\left(\frac{u(v)}{\alpha}, 1\right)$. The assumption that w_j has a sign that is the same for each type u is restrictive. Such a monotonicity restriction on one covariate will be generally needed to show separability in the variety of discrete choice models we present. The sign of w_j could be taken to be negative instead (as in the case where w_j is price), and it is trivial to extend the results to the case where w_j's sign is unknown and constant a priori.

Theorem 7.6. *Under assumptions 7.1, 7.2, 7.3, and 7.4, the distribution of utility functions in the multinomial choice model is identified with respect to the class of finite distributions.*

Proof. We verify separability. Let a finite $T \subset \Theta$ be given, where $T = \{u_1, \ldots, u_N\}$ and each u_i is a vector of utility functions. We consider I-sets of the form

$$I^T_{0,v,w} = \{u \in T \mid f((v, w), u) = 0\},$$

or just those types $u \in T$ that pick the outside good 0 at $x = (v, w)$. To show separability, we will find a $x = (v, w)$ such that $I^T_{0,v,w}$ is a singleton.

According to Definition 6.2, there exists a $v \in V$ such that $u_i(v) \neq u_k(v)$ for all $u_i \neq u_k$, $u_i, u_k \in T$. Because the set of vectors $\{u(v) \mid u \in T\}$ is finite, there exists a minimal vector $u_i(v)$. By minimal vector, we mean $u^j_k(v) > u^j_i(v)$ for some $j \in \mathcal{Y} - \{0\}$, $\forall u_k \neq u_i$. There could be multiple minimal vectors; we focus on one. Then set the vector $-w = u_i(v)$. This means that the vector of product specific utilities $u_i(v) + w = 0$ for type u_i. Now we can lower the vector w by an epsilon so that $u_i(x)$ purchases the outside good and all other types $u_j \in T - \{u_i\}$ purchase an inside good at $x = (v, w)$. Thus, $I^T_{0,v,w} = \{u_i\}$. □

7.1 Literature Review for Multinomial Choice

Matzkin (2007) surveys the literature on heterogeneous choice, emphasizing the scarcity of results on discrete choice models about the nonparametric identification of the distribution of heterogeneity, the distribution G of u, even though random coefficients are a critical tool in the empirical literature. Even papers that emphasizes the flexibility of a particular specification for heterogeneity do not formally prove identification (McFadden and Train, 2000; Rossi and Allenby, 2003; Burda et al., 2008).[9] To our knowledge, we are the first to prove the identification of general distributions of heterogeneity in multinomial choice models without relying on logit errors.

Briesch, Chintagunta and Matzkin (2009) study the identification of a discrete choice model where the payoff to choice j is, in our notation, $V(j, v_j, \omega) + w_j + \epsilon_j$, where V is a nonparametric function, w_j is a special regressor with a sign restriction, ϵ_j is an additive error and ω is a scalar unobservable that enters the utility functions for all J choices in a continuous way. Matzkin (2007) extends these results to a model where utility is $w_j + V(j, v_j, \omega)$ and ω is a vector of J unobservables, although there is no separate ϵ_j term. These specifications, with their restricted dimensions of heterogeneity, do not nest ours.

For multinomial choice, the most commonly used empirical model with unobserved heterogeneity is the random coefficients logit model. Bajari, Fox, Kim and Ryan (2009) were the first to prove the identification of the random coefficients logit model with continuous characteristics. They use calculus to show that all of the moments of the random coefficients are identified. The proof relies on linearity, $u^j(v) = v'_j\beta + \epsilon_j$ with ϵ_j having the type I extreme value distribution, but, unlike other work, only variation in v_j around its mean value is needed. Neither of the papers above deal with endogenous regressors.

Berry and Haile (2008) identify a distribution $F_t(\cdot \mid v)$ of utility *values* $t = (t_1, \ldots, t_J)$ condi-

[9]There is a some work on multinomial discrete choice models examining the nonparametric identification of the distribution of a choice-specific error ϵ_j and related parameters in models without random coefficients or random functions (Manski, 1975; Thompson, 1989; Matzkin, 1993; Lee, 1995). There is a larger literature on binary choice and ordered choice, such as Manski (1975), Cosslett (1983) and many others.

tional on v, where $t_j = u_j(v)$ for a particular v. Thus, they do not achieve full identification of a choice model where utilities are given by random utility functions or random coefficients. Identifying an unconditional distribution of utility functions rather than a conditional distribution of utility values has several uses in structural empirical work. For example, utility functions can be used to identify the utility differences of particular structural types u at old and new choice sets. For example, our theorem allows us to identify the joint distribution of $\left\{u^j(v') + w'_j - u^j(v) - w_j\right\}_{j=1}^J$, the utility improvement for each of the J products if choice sets or observable consumer characteristics in $x = (v, w)$ change. We can also identify the distribution of

$$\left\{\arg\max_{j\in\mathcal{Y}}\left\{u^j(v') + w'_j\right\} - \arg\max_{j\in\mathcal{Y}}\left\{u^j(v) + w_j\right\}\right\}, \tag{5}$$

the differences in maximized utility values, one version of a "treatment effect" for changing (v, w) to (v', w'). By contrast, the distribution $F_t(t_1, \ldots, t_J \mid v)$ does not assign utility to particular structural types, and so a researcher cannot calculate (5). The lack of utility functions prevents the researcher from computing a distribution of welfare changes, a major use of structural demand models.[10]

Chiappori and Komunjer (2009) discuss some assumptions under which they can show the identification of a multinomial choice model without additive regressors. Manski (2007) considers the identification of a counterfactual choice function when there is a fixed number of decision problems x and hence a fixed number of types with different responses at those x's. He also imposes independence between choice sets x and preferences and focuses on set identification. We point identify a distribution of utility functions on the space of all functions satisfying the WNTP, Definition 6.2.

Studying the case of $J = 1$, one inside good and one outside good, Ichimura and Thompson (1998) use the Cramér and Wold (1936) theorem for identification, which relies critically on a linear index functional form: $v'_j\beta + w_j$. We use only the quasi-linearity of $u^j(v) + w_j$ in w_j and the WNTP. A space of linear functions distinguished by the parameter β trivially satisfies the WNTP as in example 7.5. A key assumption in both papers is monotonicity in at least one regressor w_j. Ichimura and Thompson also need full support on all covariates (both v and w) to apply the Cramér-Wold theorem. Further, Ichimura and Thompson need an identification condition that reduces to our monotonicity condition that the sign of w_j in $u^j(v) + w_j$ is known. We need large support on only w in Theorem 7.6. Gautier and Kitamura (2007) provide some alternative identification arguments (the results are the same) and a computationally-simpler estimator for the model of Ichimura and Thompson. The arguments in Ichimura and Thompson and Gautier and Kitamura have not been extended to the case of multinomial choice.

[10] Using the Berry and Haile $F_t(t_1, \ldots, t_J \mid v)$, the researcher can calculate $E\left[t'_1 + w'_1 \mid x'\right] - E\left[t_1 + w_1 \mid x\right]$, as this requires only distributions of utility values at each choice set x, not the distribution of utility functions.

7.2 Purchasing Multiple Products with Complementarities or Substitutes in Preferences

None of the previous papers allow a consumer to purchase two or more goods simultaneously. In contrast, Gentzkow (2007) and Liu, Chintagunta and Zhu (forthcoming) study choice situations where each discrete choice $j = 0, \ldots, J$ indexes a bundle of composite choices. For example, a consumer can purchase cable television separately ($j = 1$), purchase an internet connection separately ($j = 2$), purchase both cable television and an internet connection together as a bundle ($j = 3$), or purchase nothing, the outside good ($j = 0$). The goal in this situation is to distinguish between explanations for observed joint purchase: are consumers observed to buy cable television and an internet connection at the same time because those who watch lots of television also have a high preference for internet service, or is there some causal utility increase from consuming both television and internet service together? The goal is to distinguish unobserved heterogeneity in preferences for products, which may be correlated across products, from true complementarities.

In our notation, unobserved heterogeneity is just captured by a distribution $G(u)$ that gives positive correlation between the utility functions $u^1(v)$, $u^2(v)$, and $u^3(v)$. True complementarities are measured by

$$\Delta(v) \equiv u^3(v) - \left(u^1(v) + u^2(v)\right).$$

If utility is $u^j(v) - w_j$ and w_j is the price of j, then $\Delta(v)$ is the monetary value of complementarities to the consumer. $\Delta(v) > 0$ represents a positive benefit from joint consumption. As utility functions are random functions across the population, there is a distribution of complementarity functions $\Delta(v)$ implied by $G(u)$.

As we have already explored in Theorem 7.6, we can identify the joint distribution of heterogeneity, which means we can identify the distribution of complementarities as a function of the joint distribution $G(u)$, if prices w_j are bundle-specific. Thus, we need to observe different choice situations where the bundle is or is not aggressively priced relative to the singleton packages. This is the data scheme for Liu et al.: they observe different bundles of telecommunications services at different prices, across geographic markets.

8 The Roy Selection Model and the Competing Risks Model

The Roy selection model is often used in labor economics to study the voluntary sorting of workers into different sectors. Workers are assumed to pick the sector that offers the highest wage, but only the wage from the chosen sector is measured in the data. In order to perform a counterfactual, one needs to identify the joint distribution of wage outcomes across all sectors, for all workers.

Borrowing notation from multinomial choice, we let $y_j = u_j(v) + \beta_j w_j$ describe the wage y_j of a worker for selecting sector j out of a set $\mathcal{J} = \{1, \ldots, J\}$ of J sectors. Wages are observed for all workers, although an outside option of non-participation can trivially be added. Compared to multinomial choice, we can identify a sector- and agent-specific parameter β_j on the special regressor w_j in each sector j because the wage y_j will be measured for the recorded sector. The transformation function

$$f(x,\theta) = \left(\arg\max_{j\in\mathcal{J}} \left\{ u^j(v) + \beta^j w_j \right\}, \max_{j\in\mathcal{J}} \left\{ u^j(v) + \beta^j w_j \right\} \right)$$

gives us both the sector chosen by each worker and the wage of the worker in the chosen sector, for $x = (v, w)$ and $\theta = (u, \beta)$ where $\beta = \left(\beta^1, \ldots, \beta^J\right)$. The goal will be to identify $G(\theta)$ or $G((u, \beta))$, the joint distribution of sector-specific return functions. Results for the Roy selection model automatically apply to the competing risks model

$$f(x,\theta) = \left(\arg\min_{j\in\mathcal{J}} \left\{ u^j(v) + \beta^j w_j \right\}, \min_{j\in\mathcal{J}} \left\{ u^j(v) + \beta^j w_j \right\} \right),$$

where we observe the disease that kills each patient and the age of death $y_j = u^j(v) + \beta^j w_j$.

Our assumptions are similar to those for the multinomial choice model.

Assumption 8.1. *Let $V \subset \mathbb{R}^K$, the support of v, be a non-empty rectangle. Let $x = (v, w) \in \mathcal{X} = V \times W_1 \times \cdots \times W_J$ where $W_j = \mathbb{R}$ for each $j \in J$.*

A type $\theta = (u, \beta)$, with $u = \left(u^1, \ldots, u^J\right)$ and $\beta = \left(\beta^1, \ldots, \beta^J\right)$, is a tuple comprising a vector of functions of the sector and individual characteristics $v \in V$ and a vector of sector-specific slope coefficients on the special regressor w_j. We will identify the distribution $G((u, \beta))$.

Assumption 8.2. *The type (u, β) is statistically independent of the observable choice set $x = (v, w)$.*

We do not discuss omitted variable bias in the Roy selection model, but results on multinomial choice in Section 9.3 extend almost automatically to this case. To show separability, we will need a monotonicity assumption for the special regressor w_j.

Assumption 8.3. *The outcome of an agent selecting sector j is $u^j(v) + \beta^j w_j$, where $\beta_j > 0$, for $j = 1, \ldots, J$.*

Assumption 8.4. *The type space Θ is a product space $\Theta_u \times \Theta_\beta$, where Θ_u is a space of functions and $\Theta_\beta \subseteq \mathbb{R}^J$. Further, each function $u^j(v)$ is real analytic, so that the induced function space for the vector of functions $\tilde{u}(v, w) = \left(\tilde{u}^1(v, w_1), \ldots, \tilde{u}^J(v, w_J)\right)$, where $\tilde{u}^j(v, w_j) = u^j(v) + \beta^j w_j$, satisfies the WNTP.*

In other words, $\tilde{u}^j(v, w_j)$ is real analytic if $u^j(v)$ is real analytic. With these assumptions, we can prove identification, and hence consistency of the minimum distance estimator.

Theorem 8.5. *Under assumptions 8.1, 8.2, 8.3, and 8.4, the distribution of sector-specific wage functions in the Roy model is identified with respect to the class of finite distributions. Also, the distribution of disease-specific death functions is identified with positive probability in the competing risks model.*

The proof and a necessary lemma are placed in an appendix. The proof directly verifies a condition we call reducibility, which is necessary and sufficient for separability.

8.1 Literature Review on the Roy Selection Model and the Competing Risks Model

Heckman and Honore (1990) study nonparametric identification in the Roy model, focusing on observed rather than unobserved heterogeneity. They present many different models. Their Theorem 12 features a wage function $g^j(v, w) + \epsilon^j$, where $g^j(v, w)$ is a homogeneous function, i.e. a function that is the same for all consumers. The error ϵ^j is a scalar unobservable, heterogeneous across the population, and independent of (v, w). In Theorem 12, non-primitive assumptions are placed on $g^j(v, w)$ so that the vector $\left(g^1(v, w), \ldots, g^J(v, w)\right)$ has range equal to $\mathbb{R}^J$. This non-primitive assumption plays a similar role as our primitive assumption of a special regressor w. Compared to Heckman and Honore (1990), we focus on the empirically relevant case where workers have unobserved heterogeneity, meaning that the functions $\left(g^1(v, w), \ldots, g^J(v, w)\right)$ vary across the units of observation. In a labor setting, if v indexes aspects of a training program offered in different intensities in different settings, some workers might respond to the training program more than others. Lee and Lewbel (2009) (see also the references therein) study the equivalent competing risks model using non-nested assumptions with Heckman and Honore (1990). They also do not allow for non-additive unobserved heterogeneity.

Note that none of the papers reviewed here use identification at infinity, which is used in Heckman (1990) to study the so-called generalized Roy model, without non-additive unobserved heterogeneity. We study the generalized Roy model with unobserved heterogeneity and without using identification at infinity in Fox and Gandhi (2009).

9 Endogenous Regressors

9.1 Endogenous Regressors and Nonadditive Random Functions

Endogenous regressors are often encountered in social-science applications. Consider the context of nonadditive random functions as in Liu (1996), where a type is a continuous mapping $g : \mathbb{R}^K \rightarrow$

$\mathbb{R}^M$. It is possible that some subset of the regressors, say the first $J < K$ regressors are not stochastically independent of an agent's type $g \in \Theta$ (due perhaps to endogenous sorting into x's). Denote the first J regressors (the endogenous regressors) as $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_J)$ and the last $N = K - J$ regressors (the exogenous regressors) as $v = (v_1, \ldots v_N)$.

Endogenous regressors show up in a variety of applications where modeling heterogeneity is critical. For example, if a type g corresponds to a demand function that is heterogeneous across markets, then characteristics such as price are often dependent with g itself (markets with less elastic demand will face a higher price). Likewise, if a type g corresponds to a production function, which is heterogeneous across firms, the firm's choice of inputs v will typically depend on the firm's technology g whenever firms choose inputs to maximize profits. For example, firms with labor-intensive technologies will hire more workers.

To handle the endogeneity problem, we extend the method of instrumental variables to allow for both heterogeneity in the primary economic equation (that is heterogeneity in random functions g), along with heterogeneity in how a type responds to the instrument. That is, we treat the instrumental variables (IV) equation as a non-additive random function as well. In particular, we assume that there exists a vector of instruments $z = (z_1, \ldots, z_J) \in Z \subset \mathbb{R}^J$ that are independent of the type g and that, along with the exogenous regressors $v \in V$, determine the endogenous regressors through an auxiliary equation $\tilde{v} = h(v, z)$.[11] A type consists of a pair of functions $\theta = (g, h)$, and the choice model in turn can be expressed as a recursive system of equations. For an economic environment $x = (v, z) \in V \times Z = \mathcal{X}$ and type (g, h), the choice model $f(x, \theta) = f((v, z), (g, h))$ predicts two outcomes, namely $\tilde{v} \in \mathbb{R}^J$ and $y \in \mathbb{R}^M$, where

$$\begin{aligned} y &= g(\tilde{v}, v) \\ \tilde{v} &= h(v, z). \end{aligned}$$

While the choice model can be solved to yield a reduced-form relationship $y = r(v, z) = g(h(v, z), v)$, the structural object of interest for policy analysis is the distribution of the causal relationship $g(\tilde{v}, v)$. In particular, if the distribution G of types (g, h) can be recovered, then we can recover the distribution of the causal or marginal effect $\frac{\partial}{\partial \tilde{v}} g(\tilde{v}, v)$, which in many cases is the main structural feature of interest.[12]

The essential feature of the model is that the exogenous variables (v, z) are stochastically

[11] We work with the just-identified case where there are as many instruments as there are endogenous regressors. Our result extends in a straightforward fashion to the overidentified case where there are more instruments than endogenous regressors.

[12] A common situation in demand estimation is that the endogenous regressor is price and price is the same across all consumers in a market. In this case, a conventional assumption is that the unobservables that are correlated with price (market-level demand shocks, say) are independent of consumer-level unobservables reflecting individual heterogeneity. Statistical independence between unobservables is a special case of our framework, which does not impose such assumptions. Further, in a world with J competing products, the unobservables reflecting demand and supply for one product may be statistically dependent with the prices of all products.

independent of the type (g, h), although the distribution of g can depend on $\tilde{v}$ conditional on the exogenous regressors (v, z), which is the source of the endogeneity problem. A special case of this model is linear two-stage least squares (2SLS) where all of the coefficients in both the outcome and IV equations are random with potential joint dependency in the coefficients across equations. That is, the random coefficients in the primary equation have an unrestricted joint distribution with the random coefficients in the IV equation. We discuss the economic importance of this example below.

We will show nonparametric identification of heterogeneity so long as the instruments satisfy a local full rank condition, which amounts to the instruments being capable of varying the endogenous regressors locally in an open set for any type. We formalize the conditions on the model below.

Assumption 9.1. *Let the support of the exogenous variables $x = (v, z)$ be the Cartesian product $V \times Z$, where $V \subseteq \mathbb{R}^N$ and $Z \subseteq \mathbb{R}^J$ are both non-empty rectangles.*

For simplicity of exposition, we take a type to be a pair (g, h) such that $g : \mathbb{R}^K \to \mathbb{R}^M$ and $h : \mathbb{R}^K \to \mathbb{R}^J$ are both vector-valued analytic functions (as defined in the appendix), which are known to satisfy the strong no ties property (at the expense of notational complexity, we could also express conditions in terms of the weak no ties property, but the generality does little to aid understanding). We also require that the IV equation h is capable of "moving around" the endogenous variables in a sense we make formal below.

Assumption 9.2. *The type space Θ consists of pairs (g, h) such that g and h are both vector-valued analytic functions where the derivative $D_z h(v, z)$ of a type's IV equation with respect to the instruments satisfies the following conditions: (i) it exists in the interior of $V \times Z$, and is continuous in $(v, z) \in V \times Z$; and (ii) the derivative $D_z h(z, v)$ with respect to z has full rank J for almost all (in the sense of Lebesgue measure) $z \in Z$.*

Such a full rank restriction is a formal way of saying that the instrument z is a locally powerful instrument almost everywhere. For any type $(g, h) \in \Theta$, almost everywhere local variation in z can induce the endogenous regressors $(\tilde{v}_1, \ldots, \tilde{v}_J)$ to vary locally in a full rank way, holding the exogenous regressors v fixed. Thus fixing $v \in V$ and for almost all $z \in Z$, the local variation in $\tilde{v}$ induced by the local variation in z is not restricted to a lower dimensional subspace.

Finally, to be valid instruments, the instruments must be independent of the agent's type.

Assumption 9.3. *The type (g, h) is stochastically independent of the instruments and exogenous regressors $x = (v, z)$.*

We now show that we can use the variation in the exogenous variables to identify the distribution G over the space of types Θ.

Theorem 9.4. *Under assumptions 9.1, 9.2, and 9.3 the distribution of nonadditive random functions (g, h) with endogenous regressors is identified with respect to the class of finite distributions.*

Proof. We proceed by showing separability of the model. Thus we take an arbitrary finite set of types $T \subset \Theta$ and seek to construct a singleton I-set. To fix I-set notation, observe that the choice variables of the model are $(y, \tilde{v}) \in \mathbb{R}^{M+J}$ and the exogenous variables are $(v, z) \in V \times Z$. Hence for any finite set of types $T \subset \Theta$, we consider I-sets that take the form

$$I^{T}_{(y,\tilde{v}),(v,z)} = \{(g,h) \in T \mid h(v,z) = \tilde{v} \,\text{and}\, g\left(h(v,z), v\right) = y\}\,.$$

Let $T_1 = \{h \mid \exists\, g \text{ such that } (g,h) \in T\}$. That is, T_1 is the set of distinct IV equations that arise within the set of types T. By definition of the WNTP (which is satisfied by analytic functions), there exists a tie breaking point $(v, z) \in V \times Z$ (which without loss can be assumed to be an interior point) such that for any distinct functions h_i and h_j in T_1, $h_i(v, z) \neq h_j(v, z)$. Consider a point $\tilde{v}^*$ from the set of values $\{h(v, z) \mid h \in T_1\}$. By construction, $\tilde{v}^*$ is attained at a unique $h \in T_1$; a unique $h \in T_1$ satisfies $\tilde{v}^* = h(v, z)$. Let us denote this unique $h \in T_1$ as h_1. By finiteness of the number of types in T_1 and the fact that each $h \in T_1$ is continuous, $h_1(t_1, t_2) \neq h(t_1, t_2)$ for all $h \in T_1$ with $h \neq h_1$ and all $(t_1, t_2) \in W \subseteq V \times Z$, where W is a sufficiently small open neighborhood containing (v, z). There are now two cases to consider.

In case 1, the set $T_2 = \{(g, h) \in T \mid h = h_1\}$ is a singleton, which contains the single type that we denote as (g_1, h_1). If we let $y^* = g_1(\tilde{v}^*, v)$, then $I^T_{(y^*,\tilde{v}^*),(v,z)}$ is a singleton, namely a set consisting of only $(g^1, h^1) \in T$.

In case 2, we have that the set $T_3 = \{g \mid (g, h_1) \in T_2\}$ is not a singleton. Observe that by Assumption 9.2, we can find a $z^* \in Z$ such that $(v, z^*) \in W$ and the Jacobian $D_z h_1(v, z^*)$ has full rank J. Furthermore, by the continuous differentiability of h_1, the Jacobian $D_z h_1(t_1, t_2)$ has full rank J for all (t_1, t_2) in a sufficiently small ball $U \subseteq W$ containing (v, z^*).

As a consequence of the Jacobian having full rank everywhere in U, the change of variable mapping $(v, z) \mapsto (h(v, z), v)$ defined over U, which we denote by R, is an open mapping by consequence of the open mapping theorem,[13] and thus the image $R(U)$ is an open set in $\mathbb{R}^K$. Now using the SNTP (which is satisfied by analytic functions), there exists $(v', z') \in U \subseteq V \times Z$ such that for all distinct functions g_i and g_j in T_3, $g_i\left(h_1\left(v', z'\right), v'\right) \neq g_j\left(h_1\left(v', z'\right), v'\right)$. We can now repeat the argument from case 1 to generate a singleton I-set. That is, we can pick any point y^* from the set of values $\{g(h_1(v', z'), v') \mid g \in T_3\} \subset \mathbb{R}^M$, and observe that by construction y^* is attained at a unique $g \in T_3$, which we can denote as g_1. Then observe the I-set $I^T_{(y^*,h_1(v',z')),(v',z')}$

[13]The matrix of partial derivatives of R is of the form $A = \begin{bmatrix} D_v h(v,z) & I_N \\ D_z h(v,z) & 0_{J,N} \end{bmatrix}$, where I_N is an identity matrix with N rows and $0_{J,N}$ is a matrix of all 0's with J rows and N columns. The matrix A is invertible because $D_z h(v, z)$ is invertible. Therefore, by the open-mapping theorem, $(v, z) \mapsto (h(v, z), v)$ is an open mapping.

is a singleton consisting of only the type (g_1, h_1). $\square$

9.2 The Generality of the Identification Result for Endogenous Regressors

The generality of the identification argument we have just proved should not be lost in the notation. A special case of Theorem 9.4 is showing identification for a linear IV model, 2SLS, with random coefficients in both the first stage and the outcome equation. Let y, $\tilde{v}$, v and z all be scalars for exposition. A type (h, g) is a system of equations

$$\begin{aligned} \tilde{v} &= a_0 + a_v v + a_z z \\ y &= b_0 + b_v v + b_{\tilde{v}} \tilde{v}, \end{aligned} \tag{6}$$

where a type θ can be represented as the unknown, random parameters $\theta = (a_0, a_v, a_z, b_0, b_v, b_{\tilde{v}})$. Theorem 9.4 shows that the joint distribution of θ, $G(\theta)$, is identified using local variation in v and z. Of course the linearity in (6) is just an example; Theorem 9.4 identifies a joint distribution over functions in a nonparametric function space.

The system (6) allows more general economic behavior than has previously been shown to be identified in the literature. In common with much of the literature, the response to $\tilde{v}$ is heterogeneous, as $b_{\tilde{v}}$ is a random coefficient. However, here the response to the instrument, a_z, is also a random coefficient. In contrast with the assumptions made in the literature on the local average treatment effect (LATE, see Imbens and Angrist (1994)) and some selection models (Vytlacil, 2002), some agents may have $a_z > 0$ and respond positively to the instrument, and other agents may have $a_z < 0$ and respond negatively to the instrument.[14] Further, the response to the instrument may be correlated with the response to the treatment. The joint distribution $G(\theta)$ may be such that those agents with the most to the gain from the treatment (a high marginal effect $b_{\tilde{v}}$) tend to have a high a_z. For a given z, this model allows agents to sort into an intensity of treatment $\tilde{v}$ based on the expected gains from treatment, $b_{\tilde{v}}$.

Consider an example. Firms differ in both their input demand functions (the first stage) and their production functions. Let y be the log output of a firm, v the age of the firm (which is independent of θ), $\tilde{v}$ the log number of workers hired by the firm (an endogenous choice variable), and z the price of labor. In this example, variation in input costs allow identification of the distribution of production functions in some industry. This framework is general. First, firms vary in how labor inputs affect outputs: the labor input elasticity $b_{\tilde{v}}$ is heterogeneous. Second, firms with higher labor elasticities may have higher input demand elasticities: $\mathrm{Corr}(a_z, b_z) > 0$.

[14] The treatment effect literature tends to focus on discrete endogenous regressors; we focus on endogenous regressors with continuous support. We show identification of the generalized Roy selection model in Fox and Gandhi (2009).

Third, there is no monotonicity in a_z, some firms may have $a_z < 0$. Say the price of labor goes up everywhere and workers are laid off at some firms. Then, due to a general equilibrium effect, some firms might actually increase their labor inputs. Identification of $G(\theta)$ allows the identification of the joint distribution of a_z and b_z as well as of the other coefficients.

9.3 Endogenous Regressors in Multinomial Choice

We now consider the endogeneity problem that arises in multinomial choice. Recalling the discussion of the multinomial choice model in Section 7, an endogeneity problem arises when an agent's preferences as captured by the utility function u are not independent of some elements of the agent's choice set (v, w). Such endogeneity could arise if, for example, the choice set (v, w) that an agent faces is partly "designed" on the basis of information related to its type or preferences u. A classic example of this source of endogeneity arises in a principal-agent relationship, in which the principal designs the menu of contracts (v, w) facing the agent using information that is correlated with the agent's type u but that is not observable by the econometrician.[15] The principal has incentives (i.e., screening) to use all information in contract design. Therefore, the endogenous choice of a menu of choices will induce a statistical endogeneity problem.[16]

In this section, we show how to solve the endogeneity problem posed by endogenous product characteristics in multinomial choice by way of a triangular system of equations that follows much the same logic as endogenous regressors in nonadditive random functions. Essentially, the triangular system jointly models the decisions of both the principal and the agent, and uses exogenous variation in the characteristics of the principal-agent relationships to achieve identification. We extend the notation from Section 7. Given $v \in \mathbb{R}^N$, $\tilde{v} \in \mathbb{R}^M$ and $w \in \mathbb{R}^J$, the agent has utility for choice j given by $u^j(\tilde{v}, v) + w_j$. Thus we let the first M elements of the vector of choice characteristics facing the agent be potentially endogenous, and denote these elements by $\tilde{v} \in \mathbb{R}^M$ and the remaining exogenous elements by $v \in \mathbb{R}^N$. We refer to these endogenous elements $\tilde{v}$ as the principal's "prices" as they are strategically set by the principal. A special case of this framework is where $N = J$ and there is one endogenous price per product.

To handle the problem, we introduce a vector of instruments $z = (z_1, \ldots, z_M) \in Z \subseteq \mathbb{R}^M$ that are stochastically independent of preferences u. In addition, the instruments are capable of shifting the endogenous choice characteristics through the principal's "pricing" or IV equation $\tilde{v} = h(v, z)$ for $z \in Z$, $v \in V$, and $h : V \times Z \to \mathbb{R}^M$.[17] Thus a type corresponds to a pair of functions $\theta = (u, h)$

[15] The standard market-level price endogeneity problem considered in footnote 12 also applies here, assuming that the price determination process can be written as the first stage of a triangular system.

[16] Pioner (2008) presents an alternative approach to identification based on a particular model of screening by a monopolist.

[17] We do not allow the w's to be endogenous or enter the pricing equation. For example, the w's could reflect variation or information that is unobserved and exogenous to seller behavior. Or the w's can capture an observable consumer attribute, such as location, that the seller cannot use as a basis for price discrimination or that does not convey information on a consumer's preferences u.

consisting of a vector-valued utility function and an IV equation. The model is such that for any economic environment $x = (v, w, z)$, the response $f((v, w, z), (u, h))$ consists of the principal's choice of prices $\tilde{v}$ and the agent's choice of product j that are linked through the recursive system

$$\begin{aligned} j &= \arg\max_{j \in \mathcal{Y}} \left\{ u^j(\tilde{v}, v) + w_j \right\} \\ \tilde{v} &= h(v, z). \end{aligned}$$

Thus a type (u, h) indexes a principal agent relationship, where the pricing equation h is potentially heterogeneous due to differing information sets or preferences among principals. Of course the joint distribution $G((u, h))$ over types allow the principal's pricing function h to be stochastically dependent with the agent's preferences u, reflecting the fact that the principal can condition its pricing policy h on information related to the agent's preferences u that is unobserved to the econometrician. The instruments z are most naturally interpreted as the marginal costs of providing each good, although they could represent any observed characteristics of the principal, including observed dimensions of its information set or any other demographic taste shifters.

By assuming exogeneity of $x = (v, w, z)$, however, we are assuming that the process that matches principals to agents is exogenous and only pricing is endogenous (otherwise agents with certain unobservable preferences may be more likely to match with principals with certain observables, thus making z an invalid instrument). Extending our framework to deal with endogenous matching is a current subject of research. Nevertheless there are numerous applied settings that fit our current version of the model. Consider Einav, Jenkins and Levin (2009), where the principal is a subprime auto dealer and the agents are the customers who exogenously arrive and desire cars with certain characteristics $(v, \tilde{v}, w)$. The principal can design contract terms $\tilde{v}$ such as the minimum down payment and the interest rate. Consumers will have heterogeneous preferences over minimum down payments and interest rates, perhaps reflecting varying liquidity constraints.

We assume that types are pairs (u, h) such that u and h are both vector-valued analytic functions, and (u, h) is independent of the regressors $(v, w, z) \in V \times \mathbb{R}^J \times Z$, where $V \subseteq \mathbb{R}^N$ and $Z \subseteq \mathbb{R}^M$ are both non-empty rectangles. Thus we exactly mimic the assumptions of the previous section (with u now playing the role of g).

Theorem 9.5. *Under Assumption 7.3, and the assumptions of Section 9.1, the distribution of* $(u, h) \in \Theta$ *in the multinomial choice model with endogenous regressors is identified with respect to the class of finite distributions.*

Proof. We provide only a sketch of the details of the proof as it is largely a repetition of techniques for showing separability that have already been illustrated in the previous theorems. For any finite set of types $T \subset \Theta$, we form a singleton I-set of the form

$$I^T_{(0,\tilde{v}),(v,w,z)} = \left\{ (u, h) \in T \mid h(v, z) = \tilde{v} \text{ and } u^j\left(h\left(v, z\right), v\right) + w_j \leq 0 \, \forall j \in \{1, \ldots J\} \right\},$$

where recall good 0 is the outside option that has a normalized utility of 0. The I-set corresponds to the set of types whose IV equation yields $\tilde{v}$ at $x = (v, w, z)$ and choose the outside good.

The proof for showing the existence of such a singleton I-set exactly follows the proof of Theorem 9.4, except with a relabelling of the relevant terms. For example, consider the last step of the proof. Instead of picking an arbitrary point $u^* \in \mathbb{R}^J$ from the set of values $\{u(h_1(v', z'), v') \mid u \in T_3\}$, we instead pick a minimal element, which by construction is attained at a unique $u \in T_3$, which we denote u_1. Then setting the special regressors w to a sufficiently small amount below $-u_1$, say w^*, then we have that $I^T_{(0,h_1(v',z')),(v',w^*,z')}$ is a singleton, consisting of only the type $(u_1, h_1) \in T$. □

9.4 Literature review on endogenous regressors

Chesher (2003) studies the nonparametric identification of a triangular system of equations where the functions in the system are non-random: the same for all types. Heterogeneity enters only through scalar error terms in each equation, and those error terms are assumed to enter the non-random functions monotonically. We allow each type to have its own function and we impose no monotonicity assumptions about how unobservables relate to outcome variables and endogenous regressors. We also do not impose monotonicity assumptions on how the instruments affect endogenous regressors, which are common in the literature on treatment effects (Imbens and Angrist, 1994; Vytlacil, 2002). Newey, Powell and Vella (1999) and Newey and Powell (2003) use a mean independence assumption in a model where heterogeneity enters the outcome equation as only an additive error, instead of a random function.[18] Imbens and Newey (2009) study a system (g, h) like ours, except that the heterogeneity in the IV equation h is restricted to be a scalar. We allow h to be a random, nonadditive function. Further, Imbens and Newey require the scalar disturbance to enter h strictly monotonically. Imbens and Newey also define to the object of interest to be what they describe as a quantile structural function. We show the full identification of all aspects of our model, namely the joint distribution of the heterogeneous functions (g, h). There are many other approaches in the nonparametric instrumental variables literature (see the above papers for more references); we know of no others that identify a distribution over systems of functions.

Hoderlein, Klemelä and Mammen (forthcoming) examine a linear triangular system such as (6), except that the coefficients a_0, a_x, a_z from the first stage are homogeneous. Only the parameters in the outcome equation are heterogeneous. Their approach relies critically on linearity, while we identify a nonparametric distribution on a nonparametric class of functions.

As discussed in Section 7.1, this is the first paper to identify a structural distribution of heterogeneity in multinomial choice that nests the standard additive random coefficients specification,

[18] Our model can be related to the nonparametric regression literature on models with additive errors, $y = m(x) + \epsilon(x)$, by the change of variables $m(x) = E_g[g(x) \mid x]$ and $\epsilon(x) = g(x) - E_g[g(x) \mid x]$. Again, many of these models do not identify an unrestricted distribution for $\epsilon(x)$ when the function ϵ is dependent with x.

$x\beta + \epsilon^j$. Therefore, we are also the first to identify a distribution of heterogeneity in such a model with endogenous regressors.

As Section 7.1 mentions, Berry and Haile (2008) do not identify a full distribution of utility functions or random coefficients, but do adopt a different approach to endogeneity. They require both individual- and market-level data and assume that the endogeneity occurs only in variables (like price) that vary at the market but not individual levels. They use individual data to trace out utility realizations within a market and variation across markets to address an endogeneity problem. One could replace their step where they trace out utility values with our Theorem 7.6. In recent work, Chiappori and Komunjer (2009) use a different approach to allowing endogeneity in multinomial choice models.

10 Marginal Effects: Is the Real Analytic Assumption Needed?

In some cases, the policy counterfactual of interest is the treatment effect associated with a marginal change in x at some point of policy relevance, $x^\star$. In this model, the economic environment $x \in \mathcal{X} \subset \mathbb{R}^m$. A type θ in the data generating process is a function $g : \mathcal{X} \to \mathbb{R}^m$, and the choice model is such that g's choice behavior at $x \in \mathcal{X}$ is $f(x, g) = g(x)$. There is some true distribution G of the functions g in the data generating process. In this section, the object of identification is the distribution H of marginal effects $Dg(x^*)$ for some specified $x^\star$, where $Dg(x)$ is the derivative of the function $g : \mathcal{X} \to \mathbb{R}^m$ at an interior point $x \in \mathcal{X}$. To repeat, we focus on identifying the distribution H of marginal effects. Focusing on this distribution centered at a point $x^\star$ will "localize" the problem and allow us to gain identification using only local variation in x, i.e. variation in x around $x^\star$. By localizing the problem, we will not need to rely on an assumption that each g is real analytic. We will assume that each g is differentiable.

Examples of marginal effects or simple transformations of marginal effects include marginal products of firms for particular inputs and demand elasticities of firms. Demand elasticities, for example, are key inputs into analyzes of competition and merger approval. Compared to models with homogeneous marginal effects, we allow each firm to have a different production function and for each firm to face a different demand curve.

The derivative of a multivariate function from $\mathbb{R}^k$ to $\mathbb{R}^m$ at a point x is a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^m$ that can be represented by the Jacobian matrix

$$Dg(x) = \begin{bmatrix} \frac{\partial g^1(x)}{\partial x_1} & \cdots & \frac{\partial g^m(x)}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g^1(x)}{\partial x_k} & \cdots & \frac{\partial g^m(x)}{\partial x_k} \end{bmatrix},$$

where $\frac{\partial g^m(x)}{\partial x_k}$ is the derivative of the mth outcome with respect to the kth input. As each type $g \in \Theta$ (assuming it is differentiable) gives rise to such a $k \times m$ Jacobian matrix $Dg(x)$ for any interior point $x^\star \in \mathcal{X}$, there exists a distribution of the Jacobian $Dg(x)$ at x^* induced by the distribution G over random functions g. Recall that the distribution of marginal effects cannot be directly observed in the data, as we observe only cross-sectional data and so cannot link the same individuals across different x environments (as can be done with panel data).

If the distribution of the marginal treatment effect is the policy counterfactual of interest, then rather than seek identification over random functions g, which is sufficient for identifying the policy counterfactual, we can seek identification of the distribution of marginal effects directly. The counterfactual of interest is the distribution over the marginal treatment effect $Dg(x)$ at an interior point $x^* \in \mathcal{X}$. Let the underlying type space Θ denote all functions from $\mathcal{X}$ to $\mathbb{R}^m$ that are differentiable at x^*. Observe that within Θ, there exist types $g \neq g'$ that differ from each other globally (there exist a $x \in \mathcal{X}$ such that $g(x) \neq g'(x)$) but have the same local behavior at x^* ($g(x^*) = g'(x^*)$ and $Dg(x^*) = Dg'(x^*)$). From a policy perspective that is concerned with the distribution of marginal effects at x^*, the distinction between g and g' is not policy relevant. In marginal effects, we will be able to proceed without invoking the WNTP (or the assumption of real analytic functions) precisely because we care only about local behavior, namely causal effects at some $x^\star \in \mathcal{X}$.

Thus we group all policy equivalent types in Θ as members of the same equivalence class. Let $\sim$ denote the equivalence relation among elements of Θ defined as $g \sim g'$ if and only if $g(x^*) = g'(x^*)$ and $Dg(x^*) = Dg'(x^*)$. The relation $\sim$ forms equivalence classes and we let the set of equivalence classes form a new type space that we denote as Θ_{x^*}. For any equivalence class $[\theta] \in \Theta_{x^*}$ (which consists of all policy identical functions from Θ), we choose any representative member function $g \in [\theta]$ to represent the choice behavior of the class. We let this representative member function g stand for the class $[\theta]$ as a whole.

The policy problem is to identify the distribution H over the policy relevant type space Θ_{x^*}. Given any rectangle $\mathcal{X} \subset \mathbb{R}^k$ containing x^*, we can show finite separability of the model and hence identification. This is a natural conclusion: given arbitrarily local variation in economic environments about x^* we can identify the distribution of marginal effects at x^*.

This is the main lemma that produces the key tie breaking result that we need to generate a singleton.

Lemma 10.1. *For any finite set of functions $g_i : \mathcal{X} \to \mathbb{R}^m$ for $i = 1, \ldots, n$ that are differentiable at $x^* \in \mathcal{X}$, if $g_i(x^\star) = g_j(x^\star)$ and $Dg_i(x^\star) \neq Dg_j(x^\star)$ for all $i \neq j$, then for any ball $B_\epsilon(x^\star)$ with $\epsilon > 0$, there exists a $x^\epsilon \in B_\epsilon(x^\star)$ such that $g_i(x^\epsilon) \neq g_j(x^\epsilon)$ for all $i \neq j$.*

Proof. To establish some notation, recall the derivative of $g : \mathcal{X} \to \mathbb{R}^m$ at $x \in \mathcal{X} \subset \mathbb{R}^k$ is a linear function that we denote $Dg[x] : \mathbb{R}^k \to \mathbb{R}^m$, and recall the value of this function at any $v \in \mathbb{R}^k$ is

$Dg[x](v)$. By assumption, $Dg_i[x^\star] - Dg_j[x^\star] \neq 0$, where 0 refers to the 0 map from $\mathbb{R}^k$ to $\mathbb{R}^m$. Then the kernel of the linear map $Dg_i[x^\star] - Dg_j[x^\star]$, which we denote by $S_{i,j}$, has dimension strictly less than k, because there exists $v \in \mathbb{R}^k$ such that $Dg_i[x^\star] - Dg_j[x^\star])(v) \neq 0$. As the finite union of subspaces $S = \cup_{i,j} S_{i,j}$ cannot equal the k-dimensional space $\mathbb{R}^k$, we can find an element $v \in \mathbb{R}^k - S$. By the construction of v, $Dg_i[x^\star](v) \neq Dg_j[x^\star](v)$ for all $i \neq j$. Hence for any positive $\lambda \in \mathbb{R}_{++}$, we have by the linearity of a derivative,

$$\frac{Dg_i[x^\star](\lambda v) - Dg_j[x^\star](\lambda v)}{||\lambda v||} = c \neq 0. \tag{7}$$

Observe that by the definition of differentiability (Carter, 2001),

$$g_i\left(x^\star + \lambda v\right) - g_j\left(x^\star + \lambda v\right) = \left(Dg_i[x^\star](\lambda v) - Dg_j[x^\star]\right)(\lambda v) + \eta(\lambda v)\,\|\lambda v\|$$

where $\eta(h) \to 0$ as $h \to 0$. Hence by (7),

$$\lim_{\lambda \to 0} \frac{g_i\left(x^\star + \lambda v\right) - g_j\left(x^\star + \lambda v\right)}{||\lambda v||} \neq 0.$$

Thus there exists $\lambda_{i,j}$ such that for all $0 < \lambda < \lambda_{i,j}$, $g_i\left(x^\star + \lambda v\right) \neq g_j\left(x^\star + \lambda v\right)$. Let $\bar{\lambda} = \min_{i,j} \lambda_{i,j}$. Then for any $B_\epsilon(x^\star)$, finding λ such that $x^\star + \lambda v \in B_\epsilon(x^\star)$ and $0 < \lambda < \bar{\lambda}$ completes the proof. □

Assumption 10.2. *$\mathcal{X}$ is a non-empty rectangle.*

Assumption 10.3. *Economic environments $x \in \mathcal{X}$ are distributed independently of types $g \in \Theta$.*

Theorem 10.4. *Under Assumptions 10.2 and 10.3, the distribution H over the type space $\Theta_{x^\star}$ is identified in the class of finite distributions. That is, the distribution of marginal effects at $x^\star$ is identified.*

Proof. The proof verifies separability of the model. Consider a finite subset of types $T = \{g_1, \ldots, g_N\} \subset \Theta_{x^*}$. An I-set is

$$I^T_{y,x} = \left\{g \in \Theta_x \mid g\left(x\right) \leq y\right\}.$$

There are two cases to consider.

The first case is that there is a unique type in T who has a minimal response at $x^\star$. Let $\{g_1(x^\star), \ldots, g_N(x^\star)\}$ be the set of responses of the types in the I-set at $x^\star$. Let $y^\star$ be a minimal vector from this set. If there a unique type g_i in T such that $y^\star = g_i\left(x^\star\right)$, then we have that $I^T_{y^*,x^\star}$ is a singleton, namely a set consisting of only the single type g_i.

The second case is when multiple types take on the minimal value $y^\star$ at $x^\star$, and thus $I^T_{y^*,x^*}$ is not a singleton set. Observe that since T is finite and since each $g \in T$ is continuous, there exists an $\epsilon > 0$, say $\bar{\epsilon}$, such that $x \in B_{\bar{\epsilon}}(x^*)$ implies that for $g \in I^T_{y^*,x}$ and $g' \in T - I^T_{y^*,x^*}$, $g'(x) \nleq g(x)$ (since by construction $g'(x^*) \nleq g(x^*)$). In addition, observe that for any pair of functions g_i and g_j in $I^T_{y^*,x^*}$, $g_i(x^\star) = g_j(x^\star)$ but $Dg_i(x^\star) \neq Dg_j(x^\star)$. Thus by Lemma 10.1, for any $\epsilon > 0$, there exists a $x^\epsilon \in B_\epsilon(x^\star)$ such that $g_i(x^\epsilon) \neq g_j(x^\epsilon)$ for all pairs of functions g_i and g_j in $I^T_{y^*,x^*}$. Choose $\epsilon > 0$ small enough so that for any $x \in B_\epsilon(x^\star)$, $x \in \mathcal{X}$ and $\epsilon < \bar{\epsilon}$. Then for any $x \in B_\epsilon(x^\star)$, there exists a minimal element y^* of the set $\{g_1(x), \ldots, g_N(x)\}$ that is attained by a unique type, $y^* = g_i(x)$ for a unique type $g_i \in T$.[19] Thus $I^T_{y^*,x}$ is a singleton consisting of only g_i. □

10.1 Literature Review for Marginal Effects

Hoderlein and Mammen (2007) and Hoderlein and Mammen (2009) (and the references in those papers) study the identification of the average (mean) marginal effect, $E[Dg(x)]$, at $x^\star$. Our framework allows us to identify the distribution H of marginal effects $Dg(x)$, not only the mean. Further, they study only the case of $m = 1$, or a scalar outcome. We allow for a vector-valued outcome variable.

Identifying a distribution of effects, not just the mean, is important when the payoff to some policy is not a linear function of the marginal effects. For example, the change in marginal product of a CEO may be leveraged over all of his subordinates, while a change in the marginal product of a production worker is not leveraged because that worker has no subordinates.

11 Conclusions

There exist few nonparametric identification theorems for the distribution of heterogeneity in many economic models estimated every day in applied microeconomics. We introduce a property of economic models, known as separability, that is a sufficient condition for identification of the distribution of heterogeneity. We also show that, under a strengthening of the definition of identifiability, that identification leads to consistency of a minimum distance estimator.

We provide new identification results, and hence new consistency results, for models of considerable applied interest. We identify distributions of utility functions in multinomial choice models, distributions of sector wage functions in the Roy model, and distributions of marginal effects in the production function model. We use instruments to correct for omitted variable bias while identifying distributions of production functions and distributions of utility functions in multinomial

[19] To see this point more precisely, observe that a minimal element of the set $\{g(x) \mid g \in I^T_{y^*,x^*}\}$ is attained by a unique type in $I^T_{y^*,x^*}$. This follows from the construction that at x, all types in $I^T_{y^*,x^*}$ make distinct choices. Let this unique type be denoted as g_i. Then by construction of $\epsilon < \bar{\epsilon}$, $y^* = g_i(x)$ continues to be a minimal element of the set $\{g(x) \mid g \in T\}$, and $g_i \in T$ is the unique type at which y^* is attained.

choice models. Our results on 2SLS extend those in the literature by allowing random coefficients in both stages of the model and for heterogeneous responses to the instrument without sign restrictions. We identify a distribution of policy changes at a particular policy level without relying on panel data or an assumption that the unknown functions are real analytic functions.

Our two main assumptions are that the distribution of heterogeneity takes on a finite number of support points (although the number of support points is learned in identification) and, in many models, that functions lie in the space of real analytic functions. The main applied use of results based on the real analytic assumption is that the identification theorems nests many semiparametric models, where the researcher specifies a parametric model known up to a finite vector of heterogeneous parameters. The distribution of heterogeneous parameters is treated nonparametrically because economic theory gives little guidance as to proper functional forms for distributions of unobserved heterogeneity.

A The Space of Real Analytic Functions Satisfies the WNTP and the SNTP

This appendix shows that the space of all real analytic functions satisfies both the WNTP and the SNTP.

Definition A.1. Let $\mathcal{X}$ be a non-empty rectangle in $\mathbb{R}^k$. A function $g : \mathcal{X} \to \mathbb{R}$ is **real analytic** if, given any interior point $\xi \in \mathcal{X}$, there is a power series in $x - \xi$ that converges to $g(x)$ for all x in some neighborhood $U \subset \mathcal{X}$ of ξ.

Real analytic functions must be infinitely differentiable.

Definition A.2. If a function $g = (g_1, \ldots, g_m) : \mathcal{X} \to \mathbb{R}^m$ is such that each of its m component functions g_i is real analytic, then g is **a vector-valued real analytic** function.

A property of the space of real analytic functions is that for any two distinct real analytic functions $g, g' : \mathcal{X} \to \mathbb{R}$, and for any open, connected set $U \subseteq \mathcal{X}$, g and g' cannot agree on the whole of U: there must exist $x \in U$ for which $g(x) \neq g'(x)$ (Krantz and Parks, 2002, Corollary 1.2.6). This property can easily be seen to extend to the space of vector-valued real analytic functions $\mathcal{A}_{\mathcal{X}}^{k,m}$. Let us call this property the *pairwise tie breaking property.* The following is now a straightforward result.

Proposition A.3. *The set of vector-valued real analytic functions satisfies the strong no ties property.*

Proof. Consider any finite set of vector-valued real analytic functions $\{g_1, \ldots, g_n\} \subset \mathcal{A}^{k,m}$. We show by induction on n that the property holds for any finite number of elements n. The base case $n = 2$ holds by the above property of vector-valued real analytic functions (for any open set $U \subseteq \mathcal{X}$, take any non-empty ball within U, which is connected, and apply the pairwise tie breaking property to this ball). Assume that the proposition holds for $n - 1$, and consider $\{g_1, \ldots, g_n\}$ and an open set $U \subseteq \mathcal{X}$, which without loss we can take to be an open ball (U contains such a ball, and balls are connected). By the induction hypothesis, there exists a point $x \in U$ such that $g_i(x) \neq g_j(x)$ for any $g_i \neq g_j$ and $i, j \in \{1, \ldots, (n-1)\}$. By the fact each g_i is continuous and the set of functions is finite, these inequalities are preserved in a small open ball $B_1 \subseteq U$ around x. Now consider the function g_n, and observe that by the pairwise tie breaking property, there exists an $x_1 \in B_1$ such that $g_n(x_1) \neq g_1(x_1)$. Furthermore, by continuity, this inequality is preserved in a small ball $B_2 \subseteq B_1$ containing x_1. Now repeat the argument, except comparing g_n with g_2, producing the a ball $B_3 \subseteq B_2$, etc. At the end of the process, a non-empty ball $B_n \subseteq B$ is produced for which any $x \in B_n$ satisfies the definition of the SNTP, i.e., $x \in B_n$ implies $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$. □

B Proof of the Identification of the Roy Model: Theorem 8.5

First we introduce a condition known as reducibility.

Definition B.1. The model $\mathcal{M}$ is **reducible** if, for any finite set of types $T \subset \Theta$, i) there exists a non-empty I-set and ii) for any non-empty I-set $I^T_{A,x}$ with two or more elements, there exists a new pair (A', x') such that the I-set $I^T_{A',x'}$ is non-empty and a strict subset of $I^T_{A,x}$.

Verifying reducibility also verifies separability.

Lemma B.2. *A model $\mathcal{M}$ is separable if and only if it is reducible.*

Proof. Say the model is reducible and let T be given. There is a non-empty I-set $I^T_{A,x}$. Then find a non-empty $I^T_{A',x'}$ that is a strict subset. Because T is finite, iteratively applying this scheme will result in a singleton I-set and hence separability. Now say the model is separable and let T be given. First, part i) of the definition of reducibility is satisfied because there exists a non-empty I-set: the singleton set from separability. For part ii), let $I^T_{A,x}$ be non-empty with at least two elements. Separability applies to any finite $\tilde{T} \subseteq \Theta$, in particular $\tilde{T} = I^T_{A,x}$. So by separability, there exists a singleton subset of $I^T_{A,x}$. □

The advantage of working with reducibility is that we can focus on an I-set with two types, without loss of generality. The difficult part is verifying part ii of Definition B.1. If a non-empty subset can be found for all I-sets with two types, then a non-empty subset can be found for all I-sets with three or more types by focusing on two of the types. We now provide a proof of the identification of the Roy model, Theorem 8.5, using reducibility.

Proof. Let a finite T be given. An I-set is

$$I^T_{j,\left[\underline{y}_j,\bar{y}_j\right],v,w}=\left\{(u,\beta)\in T\mid f\left((v,w),(u,\beta)\right)\in\{j\}\times\left[\underline{y}_j,\bar{y}_j\right]\right\}.$$

By the above argument, we can restrict attention to $T=\{(u_1,\beta_1),(u_2,\beta_2),(u_3,\beta_3)\}$ and an I-set where $I^T_{j,\left[\underline{y}_j,\bar{y}_j\right],v,w}=\{(u_1,\beta_1),(u_2,\beta_2)\}$ and so $T-I^T_{j,\left[\underline{y}_j,\bar{y}_j\right],v,w}=\{(u_3,\beta_3)\}$. We need to prove that there exists $\left(j',\left[\underline{y}'_j,\bar{y}'_j\right],v',w'\right)$ such that $I^T_{j',\left[\underline{y}'_j,\bar{y}'_j\right],v',w'}=\{(u_1,\beta_1)\}$ or $\{(u_2,\beta_2)\}$.

The first case is when $u^j_1(v)+\beta^j_1 w_j\neq u^j_2(v)+\beta^j_2 w_j$ at the initial (v,w). Then restricting the interval to $\left[\underline{y}'_j,\bar{y}'_j\right]\subset\left[\underline{y}_j,\bar{y}_j\right]$ so that $\left[\underline{y}'_j,\bar{y}'_j\right]$ includes one outcome but not the other will produce a singleton I-set. Type (u_3,β_3) will still have an outcome outside the new $\left[\underline{y}'_j,\bar{y}'_j\right]$.

The second case is when $u^j_1(v)+\beta^j_1 w_j=u^j_2(v)+\beta^j_2 w_j$ at v and w_j but either $u^j_1(\cdot)\neq u^j_2(\cdot)$ as functions or $\beta^j_1\neq\beta^j_2$. Thus, $u^j_1(v)+\beta^j_1 w_j\neq u^j_2(v)+\beta^j_2 w_j$ as functions of v and w_j. Each $u^j_1(v)+\beta^j_1 w_j$ is a real analytic function because the sum of two real analytic functions is real analytic. By a property of real analytic functions, in any open set around (v,w_j) there exists a point $\left(v',w'_j\right)$ such that $u^j_1(v')+\beta^j_1 w'_j\neq u^j_2(v')+\beta^j_2 w'_j$. By making this set small, we can keep types 1 and 2 picking choice j. By making this open set small and exploiting the continuity of $u^j_3(v)+\beta^j_3 w_j$ in (v,w_j), the outcome of type (u_3,β_3) can be kept outside the interval at $\left[\underline{y}_j,\bar{y}_j\right]$ at $\left(v',w'_j\right)$, even if type (u_3,β_3) picked sector j at $\left(j',\left[\underline{y}'_j,\bar{y}'_j\right],v',w'\right)$. Thus, we are now in the previous case and choosing $\left[\underline{y}'_j,\bar{y}'_j\right]\subset\left[\underline{y}_j,\bar{y}_j\right]$ can produce a singleton I-set.

The third case is when $u^j_1(v)+\beta^j_1 w_j=u^j_2(v)+\beta^j_2 w_j$ as functions of v and w_j and $u^k_1(v)\neq u^k_2(v)$ as functions of v for some choice k. Let $w^{\star,t}_k(v,w_j)$ be the function $\frac{1}{\beta^k_1}\left(u^j_t(v)+\beta^j_t w_j-u^k_t(v)\right)$ for type t. Type t will substitute to choice k whenever $w_k>w^{\star,t}_k(v,w_j)$. Because $u^k_1(v)\neq u^k_2(v)$ and $u^j_1(v)=u^j_2(v)$ as functions, $w^{\star,1}_k(v,w_j)\neq w^{\star,2}_k(v,w_j)$ as functions of v. Further, each $w^{\star,t}_k(v,w_j)$ is a real analytic function because it is an additive composition of real analytic functions. Exploiting the real analytic property, we can find v' such that $w^{\star,1}_k(v',w_j)\neq w^{\star,2}_k(v',w_j)$ and all discrete choices are preserved before w_k is varied. Pick $w'_k=\min\left\{w^{\star,1}_k(v',w_j),w^{\star,2}_k(v',w_j)\right\}+\epsilon$ for sufficiently small $\epsilon>0$. Then the type with the lower $w^{\star,t}_k(v,w_j)$ will pick choice k and fall out of the I-set, with the other type continuing to pick choice j with the same continuous outcome. Type 3 will at most substitute to choice k and never to choice j.

If $u^j_1(v)+\beta^j_1 w_j=u^j_2(v)+\beta^j_2 w_j$ as functions of v and w_j and $u^k_1(v)=u^k_2(v)$ for all choices $k\in J$, then, because types 1 and 2 are different types, it must be that $\beta^k_1\neq\beta^k_2$ for some choice $k\neq j$. Then $w^{\star,1}_k(v,w_j)\neq w^{\star,2}_k(v,w_j)$ as a function of v. So by the properties of real analytic functions, we can find a v' where $w^{\star,1}_k(v',w_j)\neq w^{\star,2}_k(v',w_j)$ and then repeat the argument in the previous paragraph. Thus, the model is reducible, hence separable, and hence identified. □

References

Aliprantis, Charalambos D. and Kim C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, third ed., Springer, 2006.

Allman, E.S., C. Matias, and J.A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *Ann. Statist*, 2009, *37* (6A), 3099–3132.

Bajari, Patrick, Jeremy T. Fox, Kyoo il Kim, and Stephen Ryan, "The Random Coefficients Logit Model is Identified," April 2009. NBER working paper.

Beran, R., "Prediction in random coefficient regression," *Journal of Statistical Planning and Inference*, 1995, *43* (1-2), 205–213.

_ **and P. Hall**, "Estimating coefficient distributions in random coefficient regressions," *The Annals of Statistics*, 1992, *20* (4), 1970–1984.

_ **and PW Millar**, "Minimum Distance Estimation in Random Coefficient Regression Models," *The Annals of Statistics*, 1994, *22* (4), 1976–1992.

Berry, Steven T. and Philip A. Haile, "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," 2008. Yale University working paper.

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

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

Carter, Michael, *Foundations of Mathematical Economics*, MIT Press, 2001.

Chesher, Andrew, "Identification in nonseparable models," *Econometrica*, September 2003, *71* (5), 1405–1441.

Chiappori, Pierre-André and Ivana Komunjer, "On the Nonparametric Identification of Multiple Choice Models," 2009. Columbia University working paper.

Cosslett, Stephen R., "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 1983, *51* (3), 765–782.

Cramér, H. and H. Wold, "Some Theorems on Distribution Functions," *Journal of the London Mathematical Society*, 1936, *1* (4), 290.

Einav, Liran, Mark Jenkins, and Jonathan Levin, "Contract Pricing in Consumer Credit Markets," 2009. Stanford University working paper.

Foster, A. and J. Hahn, "A consistent semiparametric estimation of the consumer surplus distribution," *Economics Letters*, 2000, *69* (3), 245–251.

Fox, Jeremy T. and Amit Gandhi, "Full Identification of the Selection Model," May 2009. University of Chicago working paper.

Gautier, Eric and Yuichi Kitamura, "Nonparametric Estimation in Random Coefficients Binary Choice Models," 2007. CREST working paper.

Gentzkow, Matthew, "Valuing New Goods in a Model with Complementarity: Online Newspapers," *The American Economic Review*, 2007, *97* (3), 713–744.

Heckman, James J., "Varieties of Selection Bias," *The American Economic Review*, 1990, *80* (2), 313–318.

_ **and Bo E. Honore**, "The Empirical Content of the Roy Model," *Econometrica*, September 1990, *58* (5), 1121–1149.

Hoderlein, Stefan and Enno Mammen, "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, September 2007, *75* (5), 1513–1518.

_ **and** _ , "Identification and estimation of local average derivatives in non-separable models without monotonicity," *Econometrics Journal*, 2009, *12*, 1–25.

_ **, Jussi Klemelä, and Enno Mammen**, "Analyzing the Random Coefficient Model Nonparametrically," *Econometric Theory*, forthcoming.

Hu, Yingyao and Matthew Shum, "Nonparametric Identification of Dynamic Models with Unobserved State Variables," June 2009. Johns Hopkins University working paper.

Ichimura, H. and TS Thompson, "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," *Journal of Econometrics*, 1998, *86* (2), 269–295.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

_ **and Whitney K. Newey**, "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, September 2009, *77* (5), 1481–1512.

Kasahara, Hiroyuki and Katsumi Shimotsu, "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 2008, *77* (1), 131–176.

Krantz, Steve G. and Harold R. Parks, *A Primer on Real Analytic Functions*, second ed., Birkhäuser, 2002.

Lee, Lung-Fei, "Semiparametric maximum likelihood estimation of polychotomous and sequential choice models," *Journal of Econometrics*, 1995, *65*, 381–428.

Lee, Sokbae and Arthur Lewbel, "Nonparametric Identification of Accelerated Failure Time Competing Risks Models," June 2009. Boston College working paper.

Lewbel, Arthur, "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 2000, *97* (1), 145–177.

Liu, Hongju, Pradeep Chintagunta, and Ting Zhu, "Complementarities and the Demand for Home Broadband Internet Services," *Marketing Science*, forthcoming. University of Connecticut working paper.

Liu, Jingou, "Minimum Distance Procedures in Nonlinear Random Coefficient Models." PhD dissertation, University of California, Berkeley 1994.

_ , "The minimum distance method in nonlinear random coefficient models," *Statistica Sinica*, 1996, *6*, 877–898.

Magnac, Thierry and Eric Maurin, "Identification and information in monotone binary models," *Journal of Econometrics*, 2007.

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

_ , "Analog estimation of econometric models," in "Handbook of Econometrics," Vol. 4, Elsevier, 1994, chapter 43, pp. 2559–2582.

_ , "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, November 2007, *48* (4), 1393–1410.

Matzkin, Rosa L., "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 1993, *58*, 137–168.

_ , "Heterogeneous Choice," in W. Newey and T. Persson, eds., *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress of the Econometric Society*, Cambridge University Press., 2007.

McFadden, D., *Conditional Logit Analysis of Qualitative Choice Behavior*, Institute of Urban and Regional Development, University of California, 1973.

McFadden, Daniel L. and Kenneth Train, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 2000, *15*, 447–470.

Newey, Whitney K. and James L. Powell, "Instrumental variable estimation of nonparametric models," *Econometrica*, September 2003, *71* (5), 1565–1578.

Newey, W.K., J.L. Powell, and F. Vella, "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 1999, *67* (3), 565–603.

Pioner, Heleno, "Semiparametric Identification of Multidimensional Screening Models," 2008. Fundaçao Getulio Vargas working paper.

Rossi, P.E. and G.M. Allenby, "Bayesian Statistics and Marketing," *Marketing Science*, 2003, *22* (3), 304–328.

Teicher, Henry, "Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 1963, *34* (4), 1265–1269.

Thompson, T.S., "Identification of Semiparametric Discrete Choice Models," 1989. Working paper, Center for Economic Research, Dept. of Economics, University of Minnesota.

Vytlacil, E., "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, January 2002, *70* (1), 331–341.

Yakowitz, S.J. and J.D. Spragins, "On the Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 1968, *39* (1), 209–214.