

Identifying Heterogeneity in Economic Choice and Selection Models Using Mixtures

Jeremy T. Fox

University of Chicago and NBER

Amit Gandhi

University of Wisconsin*

October 2008

Abstract

We show how to nonparametrically identify the distribution of heterogeneity using the linear independence of a class of structural economic choice models. We state an economic property known as reducibility and prove that reducibility ensures linear independence and hence identification. Reducibility makes verifying the identification of nonlinear models easy. We can allow for a nonparametric distribution over nonparametric functions of the data. We use our mixtures framework to prove identification in three classes of economic models: 1) continuous outcomes including triangular systems and simultaneous equations, 2) multinomial discrete choice, and 3) selection and mixed continuous-discrete choice. We rely on linear independence, not identification at infinity. For selection, we allow for essential heterogeneity in both the selection and outcome equations and fully identify the joint distribution of outcomes.

1 The Identification Problem

Heterogeneity is important in many economic problems. Some consumers may value a product characteristic more than others, so that the consumers with higher values might have less elastic demands. Firms that benefit the most from adopting a new technology might adopt it sooner, so that the returns of early adopters exceed the returns of late adopters. Workers may pick jobs in the sector where they earn the most, so that the wages of a worker who loses his preferred position may go down substantially. Drivers that buy expensive cars may be likely to use them more, so a tax on gasoline will hurt those consumers the most. Firms that enter a market may

*Thanks to Steven Durlauf, Azeem Shaikh, Morten Sørensen and Edward Vytlačil for helpful comments. Also thanks to seminar participants at Chicago, Cowles, Stanford and Wisconsin. 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 are fox@uchicago.edu and agandhi@ssc.wisc.edu.

be more productive than those that do not enter, so subsidies to entry may not raise total output much. Heterogeneity in demand functions is essential to model realistic substitution between choices (Hausman and Wise, 1978). Likewise, heterogeneity in treatment effects is essential for understanding the benefits of a particular policy intervention (Heckman, 1990).

This paper presents a general mathematical framework for establishing the identification of the distribution of heterogeneity in, possibly nonlinear, structural economic choice models. Nonparametric and flexibly parametric estimators have been proposed for estimating the distribution of heterogeneity in structural models. However, little work has been done showing the identification of such models. Without showing identification, a full proof of the consistency of nonparametric estimators cannot exist. Also, showing identification is necessary to be able to understand what types of economic model parameters can be learned from a given type of data.

A key advantage of our new mathematical framework for establishing identification is that our identification condition will be expressed in terms of an economic choice model. Thus, our condition will be relatively straightforward to verify in important classes of models where the nonparametric identification of the distribution of heterogeneity has not previously been shown. Our examples focus on generalizations of models with substantial applied use in industrial organization and labor economics: 1) the nonparametric regression model, 2) the nonparametric regression model with an endogenous regressor with a nonparametric auxiliary equation, 3) linear simultaneous equations models with all parameters being random across markets, 4) the multinomial choice demand model, including with endogenous regressors, and 5) several variants of the selection model. Heckman, Urzua and Vytlacil (2006) discuss some open questions in the selection literature. Our results on selection are worth highlighting because they address some of these issues: we allow for the selection decision to be a multinomial choice, our identification results do not rely on identification at infinity, we allow unobserved heterogeneity (random coefficients) in all equations of the model, and we identify the full joint distribution of all outcomes (not just the marginal distributions). For readers in industrial organization, we emphasize that the model of a joint brand choice and quantity decision is a special case of a selection model.

We consider a general class of economic models that can be represented as a relationship $y = g_\theta(x)$, where $y \in Y = \mathbb{R}^m$ is the response of an agent of type $\theta \in \Theta$ to the economic environment $x \in \mathcal{X} \subseteq \mathbb{R}^d$. The model is thus described by a triple $\mathcal{M} = (g, \mathcal{X}, \Theta)$, and it is informative enough to predict how any hypothetical agent $\theta \in \Theta$ would respond to any hypothetical economic environment $x \in \mathcal{X}$.

Models of the form $\mathcal{M}$ arise from economic theory. One motivating example is a random utility, multinomial choice model. There is a discrete choice set $x = \{x_1, \dots, x_J\}$ facing an agent, where each alternative $j \in \{1, \dots, J\}$ is described by a bundle of characteristics $x_j \in Z \subseteq \mathbb{R}^K$. Each agent has a type $\theta \in \Theta$ that characterizes its preferences, and thus an agent of type θ has a utility function $u_\theta(x)$ over the possible bundles $x \in Z$. Faced with the choice set x , a type θ agent chooses

the alternative $j \in \{1, \dots, J\}$ if $u_\theta(x_j) \geq u_\theta(x_k)$ for all $k \neq j$.

We do not place any structure on the type space, and in particular we allow it to be an infinite dimensional space. In the random utility model, θ could index utility functions $u_\theta(x)$ for $x \in Z$ and Θ could denote the space of real analytic functions over $Z \subseteq \mathbb{R}^K$, where no two utility functions are related by a positive monotonic transformation.¹ The only structure we will require on the type space Θ is that it be measurable.² Throughout the paper, we assume θ is statistically independent of x . Endogeneity and selection are addressed by specifying a full-information model of the endogenous regressors or selection decision.

While $\mathcal{M}$ is sufficient to predict the response of any type θ agent to any environment $x \in \mathcal{X}$, it is not sufficient to predict the aggregate response in the underlying population of types. Assuming the existence of a stable distribution of types G across economic environments $x \in \mathcal{X}$, the problem of predicting the aggregate response to x requires knowledge of the population distribution G . In the case of the random utility example, knowledge of G would allow the econometrician to predict the market demand that results from any hypothetical choice set, and the aggregate consumer welfare change that results from any hypothetical change to the choice set. Thus G is the critical ingredient from the perspective of policy analysis. The empirical problem is to identify G from the data. Our framework nests structural choice models where some or all components of θ are homogeneous: they do not randomly vary across the population.

The data consist of an i.i.d. sample of observations $\{(y_i, x_i)\}_{i=1}^N$ from the underlying population. The econometrician has no special knowledge about each unit of observation's θ_i , other than y_i and x_i . With sufficient data ($N \rightarrow \infty$), it is possible to identify $F(y | x)$, the conditional distribution function of the response, for any $y \in Y$ and $x \in X \subseteq \mathcal{X}$. Observe that the support $X \subseteq \mathcal{X}$ of economic environments in the data generating process may not be as large as the domain of economic environments admitted by the model $\mathcal{M}$. This highlights the policy importance of identifying G : it enables out of sample prediction. The question is whether G can be identified from the variation X available in the data.

We assume the distribution over types G is stable across economic environments $x \in X$. This implies that θ is stochastically independent of x in the population, allowing us to express

$$F(y | x) = G(\{\theta \in \Theta \mid g_\theta(x) \leq y\}). \quad (1)$$

Thus we immediately see that G is identified up to the measure it assigns to sets of the form $I_{y,x}^\Theta = \{\theta \in \Theta \mid g_\theta(x) \leq y\}$. The problem now becomes whether the class of sets $\mathcal{I}^\Theta = \{I_{y,x}^\Theta \mid y \in Y, x \in X\}$ is rich enough to point identify G within a class of distributions $\mathcal{G}$.

To state this problem more rigorously, let $\mathcal{F}$ denote the space of possible conditional distribution

¹We rule out functions that are positive monotonic transformations of another u_θ , $\theta \in \Theta$, because the two functions would represent the same underlying preferences.

²We will later assume that Θ is also a metrizable topological space.

functions $F(y | x)$ defined over $Y \times X$. Then we can view (1) as a mapping $L : \mathcal{G} \rightarrow \mathcal{F}$. Let F_G denote the image of G under L . We will say the model is identified relative to $\mathcal{G}$ if L is one-to-one. That is, if $G \in \mathcal{G}$, $G' \in \mathcal{G}$, $G \neq G'$, then there exists an experiment in the data $(y, x) \in Y \times X$ such that $F_G(y, x) \neq F_{G'}(y, x)$.

The identification problem can be understood as an existence problem. Identification is the problem of showing that, for any two potential distribution of types, there always exists an experiment in the data that can empirically distinguish between the distributions. In this paper, we introduce and apply a primitive condition on the economic model $\mathcal{M}$ that ensures the existence of such an experiment, and hence identification. We term this condition “reducibility” (the reason for this name will become apparent in the next section). Reducibility boils down to the following question: “For any two distinct types θ and θ' , can we find an environment $x \in X$ such that these two types have different responses, i.e., $g_\theta(x) \neq g_{\theta'}(x)$?” An affirmative answer to this question is a necessary condition for identification. If this condition fails for any two types θ and θ' , then any two distributions G and G' with support concentrated over θ and θ' could not be separately identified, because they would assign the same measure to every $I_{y,x}^\Theta \in \mathcal{I}^\Theta$. What reducibility reveals is that being able to find an environment x where two types have different responses is “almost” sufficient for identification.

We show that reducibility allows us to identify the distribution G over types in a fully nonparametric fashion. Nonparametric identification means not putting parametric structures on either the type space Θ or the distribution G . Types will be allowed to lie in an infinite dimensional space, subject only to regularity conditions, and possible distributions over types will be allowed to lie in an infinite dimensional space, subject only to regularity conditions. A lack of nonparametric identification calls into question any parametric estimator for G : apparently the parametric estimator is only consistent because of functional form restrictions on the response model $g(x, \theta)$ or parametric or finite-dimensional restrictions on the class $\mathcal{G}$, where G lives. Further, a lack of nonparametric identification means that any proposed nonparametric estimator will be inconsistent.

To achieve such generality, we must impose weak but reasonable regularity conditions on the problem. In particular, we restrict attention to a nonparametric class of distributions $\mathcal{G}$ that is almost without loss of generality as far as modeling heterogeneity in economic models. In particular, we let $\mathcal{G}$ denote the class of all multinomial distributions over Θ . By assuming that $\mathcal{G}$ equals this class, the only restriction being placed on the distribution of types is that the set of types having positive support in the population is finite. However the number of support points, the location of the support points, and their masses are completely unknown and need to be identified from the data. Thus $\mathcal{G}$ constitutes an infinite dimensional space of distributions. Furthermore, the class $\mathcal{G}$ is defined without requiring any particular structure on Θ .³

³This contrasts with the class of distributions that admit density functions, which would have to be defined contingent on the measurability properties of the underlying space Θ . This is difficult to do with general infinite-

This class of distributions is especially natural in the context of economic models because the number of individuals (and particles) in the universe is finite. The type space would ultimately be finite even if each individual had its own unique type, though the type space could be quite complicated. While arbitrarily complicated discrete distributions are sometimes inconvenient for empirical work, and hence the widespread adoption of continuous distributions in the estimation of models, a solution to the identification problem should not rely upon the abstraction that continuous measures impose upon finite populations. Because no dataset can have more than a finite number of observations and because distributions in $\mathcal{G}$ can have more support points than the atoms in the universe, it is not possible to reject the assumption on $\mathcal{G}$ with finite data. In fact, the space of multinomial distributions is dense in the space of all probability measures over Θ so long as Θ is a metrizable topological space, as most function spaces are (Aliprantis and Border, 2006, Theorem 15.10). In both an economic and mathematical sense, $\mathcal{G}$ is essentially without loss of generality.

We primarily focus on identification and not estimation. There are several approaches to the nonparametric estimation of distributions of unobserved heterogeneity. These estimators have not been unleashed on many important economic models because of the lack of nonparametric identification results; hopefully our paper will change this. Some approaches included the nonparametric maximum likelihood estimator of Laird (1978), introduced to economics in Heckman and Singer (1984). Computational approaches to approximating the NPMLE include the EM algorithm of Dempster, Laird and Rubin (1977) and the iterative procedure of Li and Barron (2000). A large literature in both frequentist and Bayesian statistics considers the estimation of finite and continuous mixtures models with and without covariates (Barbe, 1998; Day, 1969; Roueff and Rydén, 2005).⁴ Bajari, Fox, Kim and Ryan (2007) present a nonparametric, computationally simple linear least squares mixtures estimator for nonlinear models. Train (2008) considers a series of related estimators that rely on the EM algorithm for computation. Rossi, Allenby and McCulloch (2005) provide a flexible Bayesian mixtures estimator for the distribution of random coefficients in a discrete choice model. Typically, any mixtures estimator could be coupled with our identification results.

Section 2 considers general identification results for economic mixture models. Section 3 introduces an intermediate condition for economic choice models known as reducibility and shows that reducibility is sufficient for identification. After establishing our identification framework, we apply the tool of reducibility to show identification in a selection of important structural models used in applied microeconomics. Section 4 considers continuous choice models, including the nonparametric regression model, a triangular system with nonparametric and random equations, and the linear simultaneous equations model. Section 5 considers multinomial choice models and

dimensional spaces.

⁴Another use of the term “identification” in this literature is when a particular mixtures extremum estimator has a unique extremum in a finite sample (Lindsay and Roeder, 1993).

Section 6 discusses the identification of selection and discrete-continuous models. In each of the model-specific sections, we discuss how our results fit into the literature on identification for that specific class of model. Section 7 presents a fake data experiment for selection models.

2 Identification Using Mixtures

Recall the basic question is whether the class of sets $\mathcal{I}^\Theta$ generated by the model $\mathcal{M}$ is rich enough to identify G within a class of distributions $\mathcal{G}$. We now show that an affirmative answer to this question relies critically on a linear independence property of the model $\mathcal{M}$. A set of functions $\mathcal{H} = \{h(y, x; \theta) : \theta \in \Theta\}$ is linearly independent if, for any finite $T = \{\theta_i\}_{i=1}^n$, there do not exist weights $a_1, \dots, a_n$ such that $a_1 h(y, x; \theta_1) + \dots + a_n h(y, x; \theta_n) = 0$ for all $(y, x) \in Y \times X$.

To make the connection between identification and linear independence apparent, we assign to each type θ its own conditional distribution function $h(y, x; \theta) = 1[g_\theta(x) \leq y]$.⁵ Thus for each type $\theta \in \Theta$ and each $x \in X$, the function $h(y, x; \theta)$ viewed as a function of y is a Dirac delta probability distribution that takes a single step from 0 to 1 at $y = g_\theta(x)$. The advantage of expressing the underlying response function g in terms of the distribution function h is that we can now express (1) as a mixture, namely

$$F(y | x) = \int_{\Theta} h(y, x; \theta) dG(\theta). \quad (2)$$

We now adapt the theorems and proofs in Teicher (1963) and Yakowitz and Sprangins (1968) to our current setting, which yields the following necessary and sufficient condition for identification with respect to $\mathcal{G}$.

Theorem 1. *The economic model $\mathcal{M}$ is identified with respect to $\mathcal{G}$ if and only if the family of functions over $(y, x) \in Y \times X$,*

$$\mathcal{H} = \{h(y, x; \theta) : \theta \in \Theta\},$$

constitutes a linearly independent set.

Proof. For sufficiency, observe that $\mathcal{H}$ is a subset of the vector space of real valued functions over $Y \times X$. If $\mathcal{H}$ is linearly independent, then $\mathcal{H}$ is a basis for the linear span of $\mathcal{H}$.⁶ By the uniqueness of a representation by basis vectors, no two elements from $\mathcal{G}$ can yield the same $F(y | x)$ via (2).

For necessity, we prove the contrapositive: if $\mathcal{H}$ is not linearly independent, then there exists some finite $T \subseteq \Theta$, $T = \{\theta_1, \dots, \theta_n\}$ and weights $(a_1, \dots, a_n)$ such that

$$a_1 h(y, x; \theta_1) + \dots + a_n h(y, x; \theta_n) = 0 \quad (3)$$

⁵If y is a vector, we use the usual partial order on vectors in Euclidean spaces.

⁶If $\mathcal{H}$ is a set of vectors (in our case they are functions such as $h(x, y, \theta_i)$), the linear span of $\mathcal{H}$ is the set of $a_1 v_1 + \dots + a_n v_n$ for any positive integer n , any n scalars $(a_1, \dots, a_n)$, and any n vectors $\{v_1, \dots, v_n\}$.

for all $\{y, x\} \in Y \times X$.⁷ By the definition of linear independence, the weights a_i cannot all be zero. Further, recall each $h(y, x; \theta_i)$ is a cumulative distribution function, so $h(y, x; \theta_i) \geq 0 \forall y, x$. Therefore, some of the a_i 's must be negative and some must be positive. Order the a 's such that $a_i < 0$ for $i = 1, \dots, m$, where $m < n$. Then, by rearranging (3), we have that

$$\sum_{i=1}^m |a_i| h(y, x; \theta_i) = \sum_{i=m+1}^n a_i h(y, x; \theta_i) \quad (4)$$

for all pairs $\{y, x\} \in Y \times X$.

As each $h(y, x; \theta_i)$ is a cumulative distribution function, then $\lim_{y \rightarrow \infty} h(y, x; \theta_i) = 1$ for all $x \in X$ and θ_i .⁸ By setting $y = \infty$, (4) implies $\sum_{i=1}^m |a_i| = \sum_{i=m+1}^n a_i = b$, where $b > 0$ is some constant. Dividing each side of (4) by b yields

$$\sum_{i=1}^m \frac{|a_i|}{b} h(x, y; \theta_i) = \sum_{i=m+1}^n \frac{a_i}{b} h(x, y; \theta_i) \equiv F(y | x)$$

for all pairs $\{y, x\}$. Then $\sum_{i=1}^m \frac{|a_i|}{b} = \sum_{i=m+1}^n \frac{a_i}{b} = 1$. Therefore, $\left(\frac{|a_1|}{b}, \dots, \frac{|a_m|}{b}\right)$ and $\left(\frac{a_{m+1}}{b}, \dots, \frac{a_n}{b}\right)$ are both valid finite mixture distributions. The first distribution has m points of support and the second distribution has at most $n - m$ points of support. The mass points are distinct because each θ_i by construction is distinct. These two distributions use different points of support but always equal the same $F(y | x)$, as defined above, for each pair $\{y, x\}$. Therefore, data on $F(y | x)$ would not identify the mixture, so G is not identified. We have proved that linear independence is a necessary condition for identification. $\square$

As our goal is to prove the identification of the mixture distribution G in economic models, it is tempting to focus on the relatively simple proof that linear independence is a sufficient condition for identification. However, we believe the economic and statistics literatures have overlooked the beautiful insights in the proof that linear independence is also a necessary condition. Because these results are critical for our approach to verifying independence in economic models, we break them out into a separate corollary.

Corollary 1. *Let a set of functions*

$$\mathcal{H} = \{h(y, x; \theta) : \theta \in \Theta\}$$

not be linearly independent. Then there exists a finite set of types $T = \{\theta_1, \dots, \theta_n\}$, $T \subseteq \Theta$, and

⁷In this paper, we use parentheses for lists of scalars or scalar-valued functions and brackets for lists of non-scalars.

⁸If y is a vector, then this limit is taken over all elements of y to the vector that has all individual elements ∞ . If y is a discrete variable, just set y to be its maximum value.

two distributions $p = (p_1, \dots, p_n)$ and $q = (q_1, \dots, q_n)$ where

$$\sum_{i=1}^n p_i h(x, y; \theta_i) = \sum_{i=1}^n q_i h(x, y; \theta_i) \equiv F(y | x)$$

for all pairs $(y, x) \in Y \times X$ and where the two distributions have non-overlapping support: $p_i = 0$ if $q_i > 0$ and $q_i = 0$ if $p_i > 0$.

Proof. The set T and the distributions were found in the necessity proof for the previous theorem. In the notation of Theorem 1, let $p = \left(\frac{|a_1|}{b}, \dots, \frac{|a_m|}{b}, 0, \dots, 0\right)$, where there are $n - m$ zeros, and let $q = \left(0, \dots, 0, \frac{a_{m+1}}{b}, \dots, \frac{a_n}{b}\right)$, where there are at least m zeros. If any type $\theta_i \in T$ has individual weights of $p_i = q_i = 0$, simply drop θ_i from T , p_i from p and q_i from q . By the construction in Theorem 1, at least one element of q and one, distinct element of p must be nonzero. Therefore, T must have at least two elements. The distributions p and q have non-overlapping support. $\square$

2.1 Generalizations to Arbitrary Mixtures

Before continuing, we pause to allay any concern that linear independence is a mathematical condition that cannot be generalized to show the identification of arbitrary mixtures, instead of just finite mixtures with unknown support points and unknown numbers of support points, as before.⁹ Indeed, Blum and Susarla (1977) (mostly) generalize Teicher (1963) to produce a necessary and sufficient condition for the identification of arbitrary mixtures. Let Θ be a compact type space. Let $\mathcal{B}$ be the family of functions $\mathcal{B} = \{h(y, x; \theta) : (x, y) \in Y \times X\}$, where x and y are fixed instead of θ in $\mathcal{H} = \{h(y, x; \theta) : \theta \in \Theta\}$. Then Blum and Susarla prove that G in the space of arbitrary mixtures is identified if and only if $\mathcal{B}$ is dense in the space of continuous functions on Θ , in the supremum norm.¹⁰ This condition is necessary and sufficient for identification. Further, if Θ has a known, finite set of support points (which is not our assumption!), then it is relatively easy to show that if $\mathcal{H}$ is linearly independent, any function on Θ can be approximated by an element from $\mathcal{B}$. The last statement requires that $\mathcal{H}$ contains only step functions, which is the case of economic models that we focus on in this paper. So the concept of linear independence is extendable to economic models where $G \in \mathcal{G}$, the set of all distributions on θ .

We rely on Corollary 1 to show that an economic model satisfies linear independence, in the case where $G \in \mathcal{G}$. Corollary 1 arises from the proof of necessity in Theorem 1. Unfortunately, the corresponding proof of necessity in Blum and Susarla (1977) is more abstract. Instead of producing two distributions p and q with certain properties, it produces two distributions without known properties. This is because the distributions arise from the Riesz Representation theorem

⁹Arbitrary mixtures also generalize the class of distributions that admit a density.

¹⁰Bach, Plachky and Thomsen (1986) generalize Blum and Susarla. Blum and Susarla actually require that the h functions have densities; Bach et al. allow the h functions to be distributions.

applied to a linear map generated by the Hahn-Banach theorem. The proof of the Hahn-Banach theorem is, unfortunately, not constructive. Rather it relies on the Axiom of Choice. So our strategy that will use the properties of p and q from Corollary 1 is not extendable to $G \in \mathcal{G}$, although the underlying necessary and sufficient condition for identification, linear independence, does mostly generalize to the concept of $\mathcal{B}$ being dense in the space of continuous functions on Θ .

The denseness condition is harder for applied economists to verify using their own economic models. Indeed, the linear independence condition for finite mixtures is itself hard to verify, as the next section argues. Therefore, we choose to work with a simpler condition, for finite mixtures.

3 Reducibility

In the statistics literature, the family of functions $\mathcal{H}$ often corresponds to a family of distribution functions for y conditional on x parameterized by θ . In particular, the problem considered by Teicher (1963) does not include a covariate x and rather considers only a family of distribution functions for y , such as normals or gammas. For families of distribution functions, Teicher and others provide sufficient conditions based on the family of characteristic functions for the distribution functions in $\mathcal{H}$ for determining whether $\mathcal{H}$ satisfies the linear independence requirement of Theorem 1.

However, conditions on a distribution function are not applicable for our main economic setting of interest, in which $h(y, x; \theta)$ models the response of a type θ agent to a set of covariates x . A key example motivated by demand estimation is a consumer choosing from among a finite menu of J products characterized by a collection x of product-specific covariate vectors. In such settings, the distribution function h describes a Dirac delta probability measure that puts all of its mass on the index $j \in \{1, \dots, J\}$ of the alternative chosen by the agent.

Checking the condition of linear independence directly can be difficult for economic models. Consider the problem of showing that for any finite set of types $T = \{\theta_i\}_{i=1}^n \subset \Theta$, the condition $a_1 h(y, x; \theta_1) + \dots + a_n h(y, x; \theta_n) = 0$ for all pairs $(y, x) \in Y \times X$ implies that $a_i = 0$ for $i = 1, \dots, n$. This would require, for example, showing that for any such T we can find n pairs (y, x) so that the resulting matrix

$$\begin{bmatrix} h(y_1, x_1; \theta_1) & \cdots & h(y_1, x_1; \theta_n) \\ \vdots & \ddots & \vdots \\ h(y_n, x_n; \theta_1) & \cdots & h(y_n, x_n; \theta_n) \end{bmatrix}$$

has full rank. As the underlying space of types Θ is usually a continuum (a subset of $\mathbb{R}^k$ say, or a function space), it is impossible to manually check this condition for all finite subsets $T \subseteq \Theta$, as indeed there are an uncountable number of such finite sets T . Therefore, the linear independence condition is not directly useful, unless the researcher is willing to assume the true set of support

points, the set T^0 , is known in advance. However, our mathematical generality allows both the number of elements in T^0 and the elements in T^0 itself to be unknown to the researcher before identification occurs.

We now show we can use the insight stated as Corollary 1 to derive a sufficient condition for linear independence and hence identification that applied economists can straightforwardly verify for particular structural economic models $\mathcal{M}$. We now define a key concept that plays a central role in our approach to identification.

Definition 1 (I sets). *For any finite set of types $T = \{\theta_i\}_{i=1}^n \subset \Theta$, and for any $y \in \mathbb{R}^m$ and $x \in X$, the I-set $I_{y,x}^T$ is defined as*

$$I_{y,x}^T \equiv \{\theta \in T \mid g_\theta(x) \leq y\}.$$

An I-set is the set of types in some arbitrary, finite set T whose response is less than or equal to y at the covariates x . It is important that $T \subseteq \Theta$ can be any finite set, not just the true set of types. The key feature of I-sets is that they are strictly a property of the economic model $\mathcal{M}$ and variation in the data $X \subseteq \mathcal{X}$. The usefulness of I-sets lies in the fact that identifiability of G can be ensured so long as the model $\mathcal{M}$ combined with the variation in X is capable of generating sufficient variation in I-sets.

As $G \in \mathcal{G}$, we can represent G by the pair (T^0, p^0) , where the probability vector $p = (p_1^0, \dots, p_n^0) \in \Delta^{n-1}$ puts non-negative mass over a (to be identified) finite set of types $T^0 = \{\theta_i\}_{i=1}^n \subset \Theta$ for some positive integer n . Thus for $G \in \mathcal{G}$, we can express (2) as

$$F(y \mid x) = \sum_{i: \theta_i \in I_{y,x}^{T^0}} p_i^0,$$

where the summation index $i: \theta_i \in I_{y,x}^T$ means all types indices i where the corresponding $\theta_i \in T$ is also in $I_{y,x}^T$.

The first step of our identification argument is to recall Theorem 1, which established that a necessary and sufficient condition for identifiability with respect to $\mathcal{G}$ is that the model $\mathcal{M}$ combined with the variation X gives rise a family of functions $\mathcal{H} = \{h(y, x, \theta) \mid \theta \in \Theta\}$ that is linearly independent. Our question of interest is: what restrictions on the pair $(\mathcal{M}, X)$ would be sufficient to ensure the linear independence of $\mathcal{H}$? To generate such a condition, recall Corollary 1. The corollary tells us that if $\mathcal{H}$ is not linearly independent, then there exists a positive integer n along with a finite $T = \{\theta_i\}_{i=1}^n \subset \Theta$ and two distributions p and q , where

$$\sum_{i=1}^n (p_i - q_i) h(y, x; \theta_i) = 0.$$

Using I -sets, this can be restated as

$$\sum_{i=1}^n (p_i - q_i) h(y, x; \theta_i) = \sum_{i: \theta_i \in I_{y,x}^T} (p_i - q_i) = 0. \quad (5)$$

Given our construction that $p_i \neq q_i$ for every $i = 1, \dots, n$, because of the non-overlapping support in Corollary 1, (5) would generate a contradiction if the empirical model $(\mathcal{M}, X)$ is capable of producing a singleton $I_{y,x}^T$ for any finite set of types $T = \{\theta_i\}_{i=1}^n$ and some $y \in Y$ and $x \in X$. This motivates the following definition and result. Let $\mathcal{I}^T$ be the class of all I -sets for a finite set of types $T \subset \Theta$, $\mathcal{I}^T = \{I_{y,x}^T \mid y \in Y, x \in X\}$.

Definition 2 (Reducibility). *The model $(\mathcal{M}, X)$ is **reducible** if, for any finite set of types $T \subset \Theta$, there exists a class of I -sets $\mathbb{I}^T \subseteq \mathcal{I}^T$ such that*

1. *For any $I_{y,x}^T \in \mathbb{I}^T$ with at least two elements, there exists a non-empty $I_{y',x'}^T \in \mathbb{I}^T$ strictly smaller than $I_{y,x}^T$.*
2. *There exists a non-empty $I_{y,x}^T \in \mathbb{I}^T$.*

Define the process of finding a non-empty $I_{y',x'}^T$ that is strictly smaller than $I_{y,x}^T$ as reducing the I -set $I_{y,x}^T$. The following result links linear independence with reducibility.

Theorem 2. *If the model $(\mathcal{M}, X)$ is reducible, then the distribution G over Θ is identifiable with respect to $\mathcal{G}$.*

Proof. By Theorem 1, it is sufficient to show linear independence of the family $\mathcal{H}$. We prove the contrapositive. Assume that $\mathcal{H}$ is not linearly independent. By Corollary 1, there exists a finite set of types $T = \{\theta_i\}_{i=1}^n$ and distributions $(p_1, \dots, p_n)$ and $(q_1, \dots, q_n)$ over T with $p_i \neq q_i$ for every $i = 1, \dots, n$, such that for every $x \in X$ and $y \in \mathbb{R}^m$, (5) holds. Take a non-empty $I_{y,x}^T \in \mathbb{I}^T$. If $I_{y,x}^T$ is a singleton $\{\theta_i\}$, then $\sum_{i: \theta_i \in I_{y,x}^T} (p_i - q_i) = p_i - q_i \neq 0$ and both p and q cannot generate the mixture $F(y \mid x)$. If not, then reduce this I -set and start the process over, meaning check whether the I -set is a singleton. Iterate on this process. In less than n steps, the repeated application of reducibility will produce a singleton I -set and a contradiction. $\square$

It is critical to recognize that the singleton set $I_{y,x}^T$ found by reducibility is not the only pair (y, x) that is informative about the distribution G . That is, the process of reducing I -sets does not produce the experiments in the data that identify G . As the intuition behind linear independence shows, all pairs (y, x) can help distinguish the true distribution from alternatives. Said another way, the proofs of identification in Theorems 1 and 2 are not constructive. They provide no algorithm to compute the $G \in \mathcal{G}$ that satisfies (2) from knowledge of $F(y \mid x)$, although the theorems show a unique G does solve (2). The theorems are similar in spirit to a theorem that says a Nash

equilibrium in a general class of games exists and is unique, but does not provide a computer algorithm to compute the Nash equilibrium. Nevertheless, the notion of linear independence naturally suggests a practical estimator, a feature which identification at infinity strategies do not produce. We return to estimation later in the paper.¹¹

Roughly speaking, a model is reducible if for any finite set of types, the set of types who choose some alternative from a menu can be made strictly smaller but not empty by varying the alternatives in the menu, for example by making an alternative more “expensive”. We will show that reducibility is a natural and widely applicable condition. It is far from an identification at infinity condition, which is a condition about $F(y | x)$, the integral (2), at some special pair $\{y, x\}$ rather than about the individual components of Θ , which is what a condition on I -sets is about. Think about identification at infinity being about choice probabilities for the collection of all types at some special $\{y, x\}$ rather than the behavior of different types across different pairs $\{y, x\}$.¹²

This paper will be concerned with the identification of heterogeneity in a series of related economic choice models that are used often in applied microeconomics. All of our proofs for identification are based on showing reducibility, and hence linear independence and identifiability. In the definition of linear independence, we took our primitive to be the type-specific distribution function of a response conditional on covariates, $h(y, x; \theta)$. If some or all of the elements of the response $y \in Y$ are discrete, then we could have just as easily taken our primitive to be a mass function $h(j, x; \theta)$ or a mixed mass/distribution function $h(j, y, x; \theta)$ where $j \in J$ is the discrete part of the response for a finite set J and $y \in \mathbb{R}^m$ is the continuous part of the response. In terms of our identification result, these models would simply require us to adjust the definition of an I -set to suit the economic model. We study the following two additional cases.

- Discrete response: $j = g_\theta(x)$ for an integer valued j . For any finite set of types $T = \{\theta_1, \dots, \theta_n\}$, and for any $x \in X$ and $j \in J$, define the I -set

$$I_{j,x}^T = \{\theta \in T \mid j = g_\theta(x)\}.$$

- Mixed Discrete / Continuous response: $(j, y) = g_\theta(x)$ for discrete $j \in J$ and continuous $y \in \mathbb{R}^m$. For any finite set of types $T = \{\theta_1, \dots, \theta_n\}$, and for any $x \in X$, $j \in J$, and $y \in \mathbb{R}^m$, define the I -set

$$I_{j,y,x}^T = \{\theta \in T \mid j = g_{\theta,1}(x), g_{\theta,2}(x) \leq y\}.$$

¹¹We use the definition of identification in Teicher (1963). We might also want to emphasize identification on a positive measure of x . This is easy to do in most examples: if one x allows an I -set to be reduced, usually an open set (with positive probability) around that x will give the same I -set.

¹²Chamberlain (1986) coins the term identification at infinity. In generic notation, let $Q(y, x)$ be a function of two data points and let β be a parameter to be identified. The parameter β is identified at infinity if $Q(y, x) \neq \beta$ for $(y, x) < (\infty, \infty)$ but $\lim_{(y,x) \rightarrow \infty} Q(y, x) = \beta$.

3.1 Two-Step Identification with Reducibility

We now stop and consider a model that can be decomposed into two submodels. We show that a sufficient condition for the identification of the full joint distribution of all parameters in the full model is the reducibility of each submodel separately. Our two-step identification theorem is a mathematical tool that may be helpful for readers trying to apply our framework to show identification in their own models. We also use the two-step identification theorem later in this paper.

Recall the model $y = g_\theta(x)$ for $y \in Y = \mathbb{R}^M, x \in X \subseteq \mathbb{R}^K, \theta \in \Theta$. The goal is to identify the distribution G of $\theta \in \Theta$. Suppose that the type space Θ can be expressed as a product $\Theta_1 \times \Theta_2$ for $\Theta_1 = \mathbb{R}^{N_1}$. Thus an agent's type consists of a finite dimensional component and a potentially infinite dimensional component. Further suppose that the covariate space X can be expressed as a product space $X_1 \times X_2$, and $g_\theta(x) = g(x_1, \theta_1, \alpha(x_2, \theta_2))$, for $x_1 \in X_1, x_2 \in X_2$, and $\alpha : X_2 \times \Theta_2 \rightarrow \mathbb{R}^{N_2}$. Hence an agent θ 's response at $x \in X$ can be predicted on the basis of x_1 and a finite dimensional sufficient statistic $(\theta_1, \alpha(x_2, \theta_2)) \in \mathbb{R}^{N_1+N_2}$. If we held constant the value of $x_2 = c$ for some constant c , and only considered variation in the economic environment through x_1 (which is possible because X is a product space), then an agent θ can be fully described by the finite-dimensional sufficient statistic $(\theta_1, \alpha(x_2, \theta_2))$. If the economic model admits this decomposition, then we can prove identification of the underlying distribution over types $G(\theta)$ by a two step procedure, which we now establish.

Theorem 3. *The model (g, X, Θ) is identifiable with respect to $\mathcal{G}$ if the sub-models $(g, X_1, \mathbb{R}^{N_1+N_2})$ and (α, X_2, Θ_2) are both reducible.*

Proof. Observe that for each fixed value of $x_2 = c$, the underlying type distribution G induces a distribution over the the finite-dimensional sufficient statistics $(\theta_1, \alpha(x_2, \theta_2)) \in \mathbb{R}^{N_1+N_2}$. Let us denote this induced distribution as G_c , which is related to G through

$$G_c(z_1, z_2) = \Pr(\theta_1 \leq z_1, \alpha \leq z_2) = G(\{\theta \in \Theta : \theta_1 \leq z_1, \alpha(c, \theta_2) \leq z_2\}). \quad (6)$$

Observe that by the stochastic independence of θ and x , G_c is invariant to the value of x_1 .

Given any two distributions $G \neq G'$ over types θ and the reducibility of the sub-model (α, X_2, Θ_2) , we can apply Theorem 2 to show that there exists a value of x_2 , say $x_2 = c$, for which $G_c \neq G'_c$. In effect, we will show identification for a submodel where the dependent variable is a heterogeneous parameter. Consider the I -set reduction step associated with the identification of G in (6). For any finite set of types $T \subset \Theta$, let $I_{z_1, z_2, x_2}^T = \{\theta \in T : \theta_1 \leq z_1, \alpha(x_2, \theta_2) \leq z_2\}$ be a non-empty I -set with at least two elements. We need to find a (z'_1, z'_2, x'_2) that reduces this I -set. Let θ and $\tilde{\theta}$ denote two distinct types contained in I_{z_1, z_2, x_2}^T . Consider the case where the first components of the two types are not equal, $\theta_1 \neq \tilde{\theta}_1$. Suppose that θ_1 and $\tilde{\theta}_1$ differ in the k th

component with, without loss of generality, $\theta_1^k < \tilde{\theta}_1^k$. Then we have that I_{θ_1, z_2, x_2}^T reduces I_{z_1, z_2, x_2}^T : $\tilde{\theta}$ drops out of the first I -set. Next consider the case where $\theta_1 = \tilde{\theta}_1$ for every pair of distinct types θ and $\tilde{\theta}$ in I_{z_1, z_2, x_2}^T . Examine the I -set $I_{z_2, x_2}^T = \{\theta \in \Theta : \alpha(x_2, \theta_2) \leq z_2\}$. By the reducibility of the sub-model (α, X_2, Θ_2) , there exists a (z'_2, x'_2) such that $I_{z'_2, x'_2}^T$ reduces I_{z_2, x_2}^T . It follows that I_{z_1, z'_2, x'_2}^T reduces I_{z_1, z_2, x_2}^T . Thus we have reducibility of I_{z_1, z_2, x_2}^T , and we know there exists a value of x_2 , say $x_2 = c$, for which $G_c \neq G'_c$.

Thus we have established that there exists a c such that fixing $x_2 = c$, G and G' imply different distributions $G_c \neq G'_c$ over the sufficient statistic space $\mathbb{R}^{N_1+N_2}$. Furthermore, for any value of x_1 , both G_c and G'_c are each sufficient for deriving the distribution of the response $F(y \mid x_1, x_2 = c)$ implied by G and G' via

$$F(y \mid x_1, x_2 = c) = G(\{\theta \in \Theta : g_\theta(x) \leq y\}) = G_c(\{(\theta_1, \alpha) \in \mathbb{R}^{N_1+N_2} : g(x_1, \theta_1, \alpha) \leq y\}).$$

We can now use variation in only x_1 to finish the proof. In particular, since $G_c \neq G'_c$, then by reducibility of the submodel $(g, X_1, \mathbb{R}^{N_1+N_2})$, there exists $x_1 \in X_1$ for which the distribution function $F(\cdot \mid x_1, x_2 = c) \neq F'(\cdot \mid x_1, x_2 = c)$. Hence we have identification. $\square$

As we will explore later, the two step identification strategy of separately proving reducibility for each of the sub-models $(g, X_1, \mathbb{R}^{N_1+N_2})$ and (α, X_2, Θ_2) will be an convenient tool for showing the identification of selection models.

4 Continuous Response Models

We first focus attention on identifying heterogeneity in a continuous response setting, where $y = g_\theta(x)$ for $y \in \mathbb{R}^M$. A special case is a system of seemingly unrelated regression equations. Of course, a system of linear equations with random coefficients is another special case. The linear regression model with random coefficients is the simplest special case.

4.1 Systems of Nonparametric Regression Equations

We let θ be an infinite-dimensional type in order to emphasize the fully nonparametric nature of our identification results. We let each type $\theta \in \Theta$ be a function $g_\theta : X \rightarrow \mathbb{R}^m$ where $X \subset \mathbb{R}^K$ is an open set.¹³ The only restriction we impose is that each component function $g_\theta^i(x)$ for $i = 1, \dots, M$ is a real analytic function.¹⁴ A fundamental property of analytic functions that we shall exploit is

¹³This assumption rules out covariates with only discrete support. Throughout the examples, all the covariates have continuous support.

¹⁴Let X be a non-empty, open subset of $\mathbb{R}^K$. Following Abbring and van den Berg (2003), a function $g : X \rightarrow \mathbb{R}^m$ is real analytic if, given $\xi \in X$, there is a power series in $x - \xi$ that converges to $g(x)$ for all x in some neighborhood $U \subset X$ of ξ . Real analytic functions must be smooth.

the fact that for any two distinct functions $g, g' \in \Theta$, and for any open, connected set $U \subseteq X$, g and g' cannot agree on the whole of U , i.e., there exists $t \in U$ for which $g(t) \neq g'(t)$.¹⁵

Theorem 4. *The continuous response model is identified with respect to $\mathcal{G}$.*

Proof. We show reducibility of the model. For any finite set of types $T \subset \Theta$, consider any non-empty I -set $I_{y,x}^T = \{\theta \in T \mid g_\theta(x) \leq y\}$. Such a set always exists by setting y high enough. If $I_{y,x}^T$ has at least two elements, we need to show that it can be reduced. Let $Y_{y,x}^T = \{t \in \mathbb{R}^m \mid \exists \theta \in I_{y,x}^T, t = g_\theta(x)\}$ denote the set of responses for the types θ in $I_{y,x}^T$. If $Y_{y,x}^T$ contains more than two elements, say $y_i \neq y_j$, then these must differ in at least one component, say $y_i^k < y_j^k$. Thus $I_{y_i,x}^T$ would be a non-empty set that is strictly smaller than $I_{y,x}^T$, and hence the set $I_{y,x}^T$ is reducible.

On the other hand, suppose that $Y_{y,x}^T$ contains only one element. Thus all the types in $I_{y,x}^T$ have the same response y . Observe that because T is finite, we can raise y by a sufficiently small amount to y' so that both $I_{y,x}^T = I_{y',x}^T$ (no types exit or enter the I -set). This gives us that for each $\theta \in I_{y',x}^T$, $g_\theta(x) < y'$, and for each $\theta \in T \setminus I_{y',x}^T$, $g_\theta^k(x) > y'_k$ for some $k \in \{1, \dots, M\}$. By the continuity of each type's g_θ and the finiteness of T , there exists an open neighborhood $U \subset X$ containing x so that for each $x' \in U$, these inequalities continue to hold, and hence $I_{y',x'}^T = I_{y',x}^T$. Finally, as there are at least two types in $I_{y,x}^T$, say types θ_i and θ_j , we can pick an $x' \in U$ so that $g_{\theta_i}(x') \neq g_{\theta_j}(x')$. To see this, recall that $g_{\theta_i}(x) = g_{\theta_j}(x)$ by the assumption that $Y_{y,x}^T$ only contained a single element. Thus by the fact that both g_{θ_i} and g_{θ_j} are real analytic, both functions cannot agree on the whole of the open set U . Now suppose $g_{\theta_i}^k(x') < g_{\theta_j}^k(x')$ for some component $k \in \{1, \dots, M\}$. By a similar argument to the previous paragraph, we can set $y_i = g_{\theta_i}(x')$ and show that $I_{y_i,x'}^T$ is a non-empty set that reduces $I_{y,x}^T$. $\square$

The proof does not require large support in any sense for the covariates x . We only require that the variation $X \subset \mathbb{R}^K$ is an open set.

4.2 Endogenous Regressors Through a Triangular System

Endogenous regressors are often encountered in social science applications. Thus in the context of the continuous response model $g_\theta : \mathbb{R}^K \rightarrow \mathbb{R}^M$, it is possible that some subset of the regressors, say $\{x_1, \dots, x_J\} \subseteq \{x_1, \dots, x_K\}$ for $0 < J \leq K$, is not independent of the type $\theta \in \Theta$. Let us denote the endogenous regressors as $\tilde{x} = \{\tilde{x}_1, \dots, \tilde{x}_J\}$ and the exogenous regressors as $x = \{x_1, \dots, x_N\}$ for $N = K - J$. One approach to this endogeneity problem is to assume that there exists a vector of instruments $z = \{z_1, \dots, z_J\} \in \mathbb{R}^J$ that are independent of θ and that are capable of “moving

¹⁵If instead $g(t) = g'(t)$ for $t \in U$, then $f(t) \equiv g(t) - g'(t) = 0$ for $t \in U$. Proposition 1 and the second Remark on page 4 of Narasimhan (1971) shows that if a function f is 0 on U , then it is 0 everywhere. This means $g(t) = g'(t)$ for all t .

around” the endogenous regressors according to

$$\tilde{x}_i = f_\theta^i(x, z) \quad (7)$$

for $i = 1, \dots, J$.¹⁶ If we denote $f_\theta = (f_\theta^1, \dots, f_\theta^J) : \mathbb{R}^{N+J} \rightarrow \mathbb{R}^J$, then the model can be expressed as a recursive system of equations

$$\begin{aligned} y &= g_\theta(\tilde{x}, x) \\ \tilde{x} &= f_\theta(x, z) \end{aligned}$$

and a type θ thus indexes a pair of vector-valued functions $\{f_\theta, g_\theta\}$. Note that while the model can be solved to yield a reduced form $y = r_\theta(x, z)$, the structural object of interest for policy analysis is the response equation $y = g_\theta(\tilde{x}, x)$. In particular, if the distribution of types $\{f_\theta, g_\theta\}$ can be recovered, then we can recover the distribution of the causal effect $\frac{\partial}{\partial x} g_\theta(\tilde{x}, x)$, which in many cases is the main structural feature of interest. We will show nonparametric identification of heterogeneity so long as the instruments satisfy the condition of being locally valid instruments, which we now define. Again, we nest linear models with random coefficients in both the equations in the triangular system.

Let the support of the exogenous variables $\{x, z\}$ contain an open set $U \subset \mathbb{R}^{N+J}$, which we can always take to be a Cartesian product $X \times Z$ for $X \subset \mathbb{R}^N$ open and $Z \subset \mathbb{R}^J$ open. Let the type space Θ consist of all pairs $\{f_\theta, g_\theta\}$ where f_θ is real analytic, where g_θ is also real analytic and where, for each $x \in X$, the derivative matrix $Df_\theta(z, x)$ with respect to z has full rank J for each $z \in Z$. Such a full rank restriction is a formal way of saying that the instrument z is a locally valid instrument: for any type θ , local variations in z can induce the endogenous regressors $\{\tilde{x}_1, \dots, \tilde{x}_J\}$ to vary locally in a full rank way, holding the exogenous regressors x fixed. The local variation in the $\tilde{x}$ induced by the local variation in z is not restricted to a lower dimensional subspace. 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 5. *The IV model is identified with respect to $\mathcal{G}$.*

Proof. The endogenous variables of the model are $\{y, \tilde{x}\} \in \mathbb{R}^{M+J}$ and the exogenous variables are $\{x, z\} \in X \times Z$. Hence for any finite set of types $T \subset \Theta$, an I -set takes the form

$$I_{y, \tilde{x}, x, z}^T = \{\theta \in \Theta \mid f_\theta(x, z) \leq \tilde{x} \text{ and } g_\theta(f_\theta(x, z), x) \leq y\}.$$

We proceed by showing reducibility for any such non-empty I -set with at least two elements. First,

¹⁶We 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.

given the continuity of the system $\{f_\theta, g_\theta\}$ for any $\theta \in \Theta$, we can mimic the first part of the proof of Theorem 4. Without loss of generality assume that $I_{y, \tilde{x}, x', z'}^T = I_{y, \tilde{x}, x, z}^T$ for all $\{x', z'\} \in V$ where V is an open, connected set containing $\{x, z\}$. Let θ_i and θ_j denote two types in $I_{y, \tilde{x}, x, z}^T$. There are two cases of interest.

The first case to consider is where $f_{\theta_i} \neq f_{\theta_j}$. Because both functions are analytic, there exists $\{x', z'\} \in V$ for which $f_{\theta_i}(x', z') \neq f_{\theta_j}(x', z')$. Suppose without loss of generality, $f_{\theta_i}^k(x', z') < f_{\theta_j}^k(x', z')$ for some $k \in \{1, \dots, J\}$. By letting $\tilde{x}^i = f_{\theta_i}(x', z')$, a similar argument as in the proof of Theorem 4 shows that $I_{y, \tilde{x}^i, x', z'}^T$ reduces $I_{y, \tilde{x}, x, z}^T$.

The second case is where $f_{\theta_i} = f_{\theta_j} = f$ and $g_{\theta_i} \neq g_{\theta_j}$. Because the derivative matrix $D_z f_\theta(z, x)$ with respect to z has full rank J for each $z \in Z$, the change of variable mapping $(x, z) \mapsto (f(x, z), x)$ (which we can denote as L) is an open mapping, by a consequence of the open mapping theorem.¹⁷ Combined with the fact that both g_{θ_i} and g_{θ_j} are analytic, this implies that there exists $(x', z') \in V$ such that $g_{\theta_i}(f(x', z'), x') \neq g_{\theta_j}(f(x', z'), x')$ (this follows more precisely because $L(V)$ is an open, connected set in $\mathbb{R}^{J+N}$). Then just like the previous case, suppose without loss of generality that $g_{\theta_i}^k(f(x', z'), x') < g_{\theta_j}^k(f(x', z'), x')$ for some component $k \in \{1, \dots, M\}$. Letting $y_i = g_{\theta_i}(f(x', z'), x')$, we have that $I_{y_i, \tilde{x}, x', z'}^T$ reduces $I_{y, \tilde{x}, x, z}^T$. $\square$

4.3 Linear Simultaneous Equations

Simultaneous equations arise when two groups of agents, say firms and consumers, interact to create the dependent variables. The standard example in economics is that suppliers and demanders interact in market equilibrium, where supply equals demand. Let the model be

$$\begin{aligned} q_i^d &= \alpha_{0,i} + \alpha_{1,i}x_1^d + \dots + \alpha_{K,i}x_K^d + \gamma_i^d p_i \\ q_i^s &= \beta_{0,i} + \beta_{1,i}x_1^s + \dots + \beta_{L,i}x_L^s + \gamma_i^s p_i \\ q_i^d &= q_i^s, \end{aligned}$$

where i can be thought of as a market, q_i^d is the quantity demanded in market i at price p_i , q_i^s is the quantity supplied in market i at price p_i , and in equilibrium supply equals demand, so $q_i^d = q_i^s$. The K demand shifters are $x_1^d, \dots, x_K^d$ and the L supply shifters are $x_1^s, \dots, x_L^s$. A type θ_i of a market is

$$\theta_i = (\alpha_{0,i}, \alpha_{1,i}, \dots, \alpha_{K,i}, \gamma_i^d, \beta_{0,i}, \beta_{1,i}, \dots, \beta_{L,i}, \gamma_i^s).$$

Our goal is to estimate the distribution of types, $G(\theta)$. This of course includes the distribution of the demand and supply curve slopes, γ_i^d and γ_i^s . Under standard economic assumptions, $\gamma_i^d < 0$

¹⁷The matrix of partial derivatives of L is of the form $A = \begin{bmatrix} D_x f_\theta(z, x) & I_N \\ D_z f_\theta(z, x) & 0_J \end{bmatrix}$, where I_N is an identity matrix with N rows and 0_J is a square matrix of all 0's with J rows. The matrix A is invertible because $D_z f_\theta(z, x)$ is invertible. Therefore, by the open mapping theorem, $(x, z) \mapsto (f(x, z), x)$ is an open mapping.

and $\gamma_i^s > 0$: consumers buy more when the price decreases and suppliers produce more when the price increases.

The simultaneous equations model suffers from an endogeneity problem because the market clearing price and quantity pair (p_i, q_i) is a function of all the random coefficients in θ_i . To see this, solve for the reduced forms of price and quantity

$$\begin{aligned} p_i(x^d, x^s, \theta_i) &= \frac{1}{\gamma_i^s - \gamma_i^d} (\alpha_{0,i} + \alpha_{1,i}x_1^d + \dots + \alpha_{K,i}x_K^d) - \frac{1}{\gamma_i^s - \gamma_i^d} (\beta_{0,i} + \beta_{1,i}x_1^s + \dots + \beta_{L,i}x_L^s) \\ q_i(x^d, x^s, \theta_i) &= \frac{\gamma_i^s}{\gamma_i^s - \gamma_i^d} (\alpha_{0,i} + \alpha_{1,i}x_1^d + \dots + \alpha_{K,i}x_K^d) - \frac{\gamma_i^d}{\gamma_i^s - \gamma_i^d} (\beta_{0,i} + \beta_{1,i}x_1^s + \dots + \beta_{L,i}x_L^s). \end{aligned}$$

Note that this is a generalization of the standard linear simultaneous equations model, which worries about only endogeneity from the intercepts, $\alpha_{0,i}$ and $\beta_{0,i}$. In a duplication of notation, the reduced forms can be rewritten as

$$\begin{aligned} p_i(x^d, x^s, \pi_i) &= \pi_{0,i}^p + \pi_{0,i}^{p,d}x_1^d + \dots + \pi_{K,i}^{p,d}x_K^d + \pi_{1,i}^{p,s}x_1^s + \dots + \pi_{L,i}^{p,s}x_L^s \\ q_i(x^d, x^s, \pi_i) &= \pi_{0,i}^q + \pi_{0,i}^{q,d}x_1^d + \dots + \pi_{K,i}^{q,d}x_K^d + \pi_{1,i}^{q,s}x_1^s + \dots + \pi_{L,i}^{q,s}x_L^s. \end{aligned}$$

The reduced form parameters $\pi_i = (\pi_i^p, \pi_i^s)$ in terms of the structural parameters θ_i are

$$\begin{aligned} \pi_i^p &= \left\{ \frac{1}{\gamma_i^s - \gamma_i^d} \alpha_{0,i} - \frac{1}{\gamma_i^s - \gamma_i^d} \beta_{0,i}, \frac{1}{\gamma_i^s - \gamma_i^d} \alpha_{1,i}, \dots, \frac{1}{\gamma_i^s - \gamma_i^d} \alpha_{K,i}, \frac{-1}{\gamma_i^s - \gamma_i^d} \beta_{1,i}, \dots, \frac{-1}{\gamma_i^s - \gamma_i^d} \beta_{L,i} \right\} \\ \pi_i^q &= \left\{ \frac{\gamma_i^s}{\gamma_i^s - \gamma_i^d} \alpha_{0,i} - \frac{\gamma_i^d}{\gamma_i^s - \gamma_i^d} \beta_{0,i}, \frac{\gamma_i^s}{\gamma_i^s - \gamma_i^d} \alpha_{1,i}, \dots, \frac{\gamma_i^s}{\gamma_i^s - \gamma_i^d} \alpha_{K,i}, \frac{-\gamma_i^d}{\gamma_i^s - \gamma_i^d} \beta_{1,i}, \dots, \frac{-\gamma_i^d}{\gamma_i^s - \gamma_i^d} \beta_{L,i} \right\}. \end{aligned}$$

All of this notation is only valid when the demand shifters x^d and the supply shifters x^s do not overlap. If some variable k is both a demand and a supply shifter, then we have only the composite reduced form coefficients for variable k ,

$$\pi_{k,i}^p = \frac{1}{\gamma_i^s - \gamma_i^d} \alpha_{k,i} - \frac{1}{\gamma_i^s - \gamma_i^d} \beta_{0,i}, \text{ and } \pi_{k,i}^q = \frac{\gamma_i^s}{\gamma_i^s - \gamma_i^d} \alpha_{k,i} - \frac{\gamma_i^d}{\gamma_i^s - \gamma_i^d} \beta_{0,i}.$$

Let $\pi_i \in \Pi \subseteq \mathbb{R}^{1+K+L}$. First we identify the distribution $G_\pi(\pi) \in \mathcal{G}_\pi$ for the reduced form parameters. For an arbitrary finite set $T \subseteq \Theta$, let the I -set be

$$I_{p,q,x^d,x^s}^T = \{\theta_i \in T \mid p_i(x^d, x^s, \pi_i) \leq p, q_i(x^d, x^s, \pi_i) \leq q\}.$$

To focus on simultaneity and not omitted variable bias, we maintain θ_i is independent of (x^d, x^s) , the vectors of demand and supply shifters. Under this independence assumption, the joint distribution of the price reduced form parameters, $G_\pi(\pi)$, is identified by an appeal to the seemingly

unrelated regressions theorem, Theorem 4. The application of Theorem 4 assumes that the support of the vector (x^d, x^s) is an open rectangle in $\mathbb{R}^{K+L}$.

Take a given $\pi_i \in \Pi$. The corresponding structural parameter θ_i is identified under the well-known rank and order conditions. For our case of only one endogenous variable in each structural equation, the order condition is that at least one x_k^d is excluded from the supply equation (x_k^d is not in the vector of supply shifters x^s) and, similarly, that one x_l^s is not in x^d . Because x_k^d must be included in x^d , this means the corresponding coefficient $\alpha_{k,i}$ can never be zero, unless there is overidentifying information in other demand shifters excluded from supply. If these conditions are met, the distribution $G(\theta_i)$ of the structural parameters can be recovered from the distribution of the reduced form parameters, $G(\pi_i)$, by a pointwise change of variables. This discussion is summarized in the following theorem.

Theorem 6. *Consider a parameter space Θ for the supply and demand model, where each $\theta_i \in \Theta$ satisfies*

$$\gamma_i^d \leq 0, \gamma_i^s \geq 0, \alpha_{i,1} \neq 0, \dots, \alpha_{i,K} \neq 0, \beta_{i,1} \neq 0, \dots, \beta_{i,K} \neq 0.$$

Further, at least one element of x^d is excluded from x^s and at least one element of x^s is excluded from x^d . Then if the support of the vector (x^d, x^s) is an open rectangle in $\mathbb{R}^{K+L}$, the reduced form distribution $G_\pi(\pi) \in \mathcal{G}_\pi$ and the structural distribution $G(\theta) \in \mathcal{G}$ are both identified.

The conditions that all demand and supply shifters never have zero coefficients is a bit strong; identification really requires that, for each θ_i , there is one shifter included in demand and excluded from supply and one shifter included in supply and excluded from demand. The restriction that demand curves slope downwards and supply curves slope upwards could be weakened to $\gamma_i^d \neq -\gamma_i^s$ for all $\theta_i \in \Theta$. These assumptions ensure that there is a unique structural parameter θ that generates each reduced form π .

We omit the proof because the previous discussion, an appeal to Theorem 4 to identify the reduced form and then an appeal to standard simultaneous identification results pointwise, is sufficient. The appeal to our reducibility condition is nested in the reference to Theorem 4, but in the background reducibility is being used to show linear independence and hence identification of the mixture distributions for both π and θ . Our arguments can easily be extended to systems of more than two equations by formalizing the usual order and rank conditions found in econometrics textbooks.

4.4 Literature Review for Continuous Outcomes

A large literature focuses on the nonparametric identification of the distribution of random coefficients in only the linear regression model (Beran and Millar, 1994; Hoderlein, Klemelä and Mammen, 2008). To our knowledge, there is no general treatment of the identification of heteroge-

neous coefficients in parametric, nonlinear models. We go beyond even this and show identification where a particular type represents a real analytic function. Further, we allow for endogenous regressors in a triangular system. Indeed, all of our results allow for systems of equations. We know of no other work that comes close to identifying a nonparametric distribution over an infinite-dimensional, nonparametric class of functions. Thus, we dramatically extend the results from the previous literature.

5 Discrete Choice over Differentiated Products

Discrete choice over differentiated products is a key model used in empirical industrial organization to model consumer demand. Demand functions are useful for measuring market power when combined with a supply model. Demand functions can also be used to predict the welfare gain from new goods, among other uses. This section shows how discrete choice models of differentiated products demand are nonparametrically identified within our framework. The differentiated products framework is popular because it allows products to be distinguished by a parsimonious list of product characteristics. Consumers mostly have preferences over these product characteristics rather than individual products themselves. This allows the researcher to predict demand and measure welfare when characteristics change.

We first consider the case in which there is no price endogeneity; the main driver of identification is variation in choice sets. We then introduce price endogeneity, where identification requires a source of both within and across market variation in choice sets.

5.1 Discrete Choice Models Without Price Endogeneity

Let the utility from type θ purchasing product j be $U_{\theta,j} = u_{\theta}^j(u_{\theta}^j(x_j), w_j)$. Both the outer function $v_{\theta}^j(\cdot, \cdot)$ and choice- j and type- θ The outer function $v_{\theta}^j(\cdot, \cdot)$ is monotone its second, scalar argument $w_j \in \mathbb{R}^+$. Furthermore,

One example is that w_j could be the price of good j . In this case, $u_{\theta,j}$ is type θ 's reservation price for product j , and it would be more natural to write $u_{\theta,j} - w_j$. However, w_j could be some non-price covariate or, with individual data, an interaction of a consumer and product characteristic. We shall refer to each product's w_j as product j 's regressor. A typical assumption in the demand estimation literature is that a consumer's reservation price for a product j is driven by a vector of underlying product characteristics $x_j = (x_{j,1}, \dots, x_{j,K}) \in \mathbb{R}^K$, and preferences are such that $u_{\theta,j} = u_{\theta}^j(x_j)$, where $\theta \in \Theta$ denotes a type. A type $\theta \in \Theta$ indexes a J -tuple of such sub-utility functions $u_{\theta}(x_1, \dots, x_J) = (u_{\theta}^1(x_1), \dots, u_{\theta}^J(x_J))$.¹⁸

¹⁸Allowing each product j to have its own utility function over product attributes reflects an interaction between a product's indicator variable and the product's attributes in consumer preferences. The more-usual empirical specification where the functions $(u_{\theta}^1(x), \dots, u_{\theta}^J(x))$ are the same, up to a consumer-and-product-specific error

Implicit in the quasilinear representation of preferences $U_{\theta,j}$ 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 $u_{\theta,j} + w_j$ by a positive constant. Thus if w admitted a type specific coefficient α_θ , then the type $\{u_\theta, \alpha_\theta\}$ would have the exact same preferences as the type $\left\{\frac{u_\theta}{\alpha_\theta}, 1\right\}$. The assumption that w_j has a sign that is the same for each type θ is restrictive. Such a monotonicity restriction on one covariate only is needed to show reducibility in discrete outcome models. The sign of w_j could be assumed to be negative instead (think of the example where w_j is price), but we will work with a positive sign on w_j . It is trivial to extend the results to the case where w_j 's sign is unknown. We also impose a location normalization by introducing an outside good $j = 0$ for which $U_{\theta,0} = 0$ for all types θ .

Assume that the J -tuple of product attributes $\{x_1, \dots, x_J\}$ varies over a set $X \subset \mathbb{R}^{JK}$ that contains an open, connected subset. We can assume for simplicity that X itself is open and connected. Assume also that the J -tuple of regressors $\{w_1, \dots, w_J\}$ varies over $W = \mathbb{R}^J$ and that the entire of menu of products $(\{x_1, \dots, x_J\}, \{w_1, \dots, w_J\})$ varies over the product set $X \times \mathbb{R}^J$. Faced with any such menu $(\{x_1, \dots, x_J\}, \{w_1, \dots, w_J\})$, a type $\theta \in \Theta$ chooses product j when $U_{\theta,j} \geq U_{\theta,k}$ for all $k = 0, \dots, J$. The econometrician can identify conditional market shares $f(j \mid (\{x_1, \dots, x_J\}, \{w_1, \dots, w_J\}))$ for $j = 0, \dots, J$. Our goal is to identify the distribution of the types $G(\theta)$, where different types θ index different sub-utility functions u_θ .

Our strategy for proving identification works in two stages. Holding fixed the menu of product characteristics $x = \{x_j\}_{j=1}^J$, we use variation in the regressors $\{w_1, \dots, w_J\}$ to identify the joint distribution of values $u = (u^1, \dots, u^J) = (u_\theta^1(x), \dots, u_\theta^J(x)) \in \mathbb{R}^J$ in the population for each such $x \in X$. In the second stage, after the distribution of u_θ has been recovered for each menu of product characteristics $x \in X$, the resulting joint distribution $F(u_\theta \mid x)$ can be decomposed using the continuous response model to recover the distribution over types $\theta \in \Theta$.

In the first stage, a type chooses the product that maximizes the type's utility. For any finite set of types, we can define the I -set

$$I_{j,w}^T = \left\{ \theta \in T \mid u_\theta^j + w_j \geq u_\theta^k + w_k \forall k = 0, \dots, J, k \neq j \right\},$$

for any $j = 0, \dots, J$ and $w \in W$. We adopt the convention that the outside good is such that $u_\theta^0 = 0$ and $w_0 = 0$ for all types θ .

Theorem 7. *For any $x \in X$, the joint distribution $G(u^1, \dots, u^J \mid x)$ of subutilities is identified.*

Proof. We proceed by reducibility. Consider any non-empty I -set associated with the outside good. We can always ensure a non-empty set by decreasing $w \in \mathbb{R}^J$. Consider any two types $\theta, \theta' \in T$, $u_\theta \neq u_{\theta'}$. Suppose in particular $u_\theta^j > u_{\theta'}^j$, for some product j . Then we can raise w_j so that

term $\epsilon_{\theta,j}$, is a special case of our framework.

$u_\theta^j + w_j \geq 0$ but $u_{\theta'}^j + w_j < 0$. Hence type θ drops out of the I -set but θ' remains. We have reducibility. $\square$

Thus for any $x \in X$, we have identified the conditional, joint distribution of the subutilities, $G(u^1, \dots, u^J \mid x)$. As $u^j = u_\theta^j(x_j)$, we can use variation in $x \in X$ to identify the distribution over types $G(\theta)$, where $\theta \in \Theta$ indexes a J -tuple of sub-utility functions $u_\theta = (u_\theta^1(x), \dots, u_\theta^J(x))$. If X is open and connected and Θ consists of the set of subutility functions that are real analytic, then we have the following result.

Theorem 8. *The distribution $G(\theta)$ over the type space Θ consisting of all real analytic subutility functions u_θ is identified.*

Proof. Observe that the utility model $u^j = u_\theta^j(x_j)$ for $j = 1, \dots, J$ is a special case of the continuous response model $u^j = u_\theta^j(x)$ for $x = (x_1, \dots, x_J) \in X$, where X is an open and connected subset of $\mathbb{R}^{JK}$. As the distribution function over subutilities $G(u^1, \dots, u^J \mid x)$ is identified for each $x \in X$, we can apply Theorem 4 to show that the distribution G over Θ is identified. Theorem 4 requires that each $u_\theta : X \rightarrow \mathbb{R}^J$ be real analytic. $\square$

5.2 Price Endogeneity

In independent work, Berry and Haile (2007) use a two-step approach for the identification of multinomial choice models. Their approach relies on a separation between within-market and across-market variation. A key contribution is that they allow for one unobserved product characteristic $\xi_{j,m}$ for each product j in market m . In their first stage, they rely on within-market (so $\xi_{j,m}$ is fixed) variation in w to identify the marginal $G^j(u^j \mid m)$, where m is the market and the market-specific characteristics $\{x_{j,m}, \xi_{j,m}\}_{j=1}^J$ are thus held fixed. In a second stage, for each product j in market m they compute

$$\delta_{j,m} = \text{median}[u^j \mid m].$$

Assuming that u^j is strictly increasing in $\xi_{j,m}$, they then perform a Chernozhukov and Hansen (2005) nonparametric, instrumental variables, quantile regression to identify $\xi_{j,m}$ for each product j in each market m . In a third stage, the joint distribution of sub-utilities $G(u^1, \dots, u^J)$ conditional on market-level product characteristics $\{x_{j,m}, \xi_{j,m}\}_{j=1}^J$, their primitive, is identified.

We can likewise employ a similar strategy to handle price endogeneity. In particular, the nonparametric regression required to recover $\xi_{j,m}$ for each product j in each market m in the support of the data generating process would occur as an intermediate step between our first (Theorem 7) and second (Theorem 8) stages of identification. Our identification arguments would identify the joint distribution of the subutilities in a market m , $G(u^1, \dots, u^J \mid m)$, in the first stage.

5.3 Literature Review for Discrete 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 θ , even though random coefficients are a critical tool in the empirical literature. Even papers that emphasize the flexibility of a particular specification for heterogeneity do not formally prove identification (McFadden and Train, 2000; Rossi and Allenby, 2003).¹⁹

Briesch, Chintagunta and Matzkin (2007) study the identification of a discrete choice model where the payoff to choice j is $V(j, s, x_j, \omega) + \epsilon_j$, where V is a nonparametric function and ω is a scalar unobservable that enters the utility functions for all J choices. They use a Lewbel (2000) special regressor argument and make assumptions on the distributions functions of the errors so that a theorem from Teicher (1961) can be used.²⁰ For multinomial choice, the most commonly-used empirical model with unobserved heterogeneity is the random coefficients logit model. Bajari, Fox, Kim and Ryan (2007) were the first to prove the identification of the random coefficients logit model with continuous characteristics using calculus to show that all of the moments of the random coefficients are identified. The proof relies on linearity, $u_{\theta,j} = x'_j \beta_\theta$, but, unlike other work, only variation in $x'_j \beta_\theta$ around the value $x'_j \beta_\theta = 0$ is needed. None of the papers in this subsection deal with endogenous regressors.

Berry and Haile (2007) and our paper simultaneously developed approaches to identifying the distribution of heterogeneity in multinomial choice models.²¹ In their Section 4, Berry and Haile adopt an identification-at-infinity strategy, where the payoffs of all but $J - 2$ of the choices are set to minus infinity, in order to reduce one step of identification to a binary choice problem. Recognizing identification at infinity is empirically implausible, in Section 5 they adopt a quantile identification strategy to partial identification, where, for binary choice, they identify the τ th quantile of the utility for choice j if they can find covariates that push the choice probability of choice j to be τ . If τ can be 1, they get full identification. We do not discuss any approaches that rely on identification at infinity. While the necessary and sufficient condition for identification, linear independence or its generalizations, has nothing to do with pushing the probability of any choice to be 1, our sufficient condition for linear independence and hence identification, reducibility, does imply that high values of w_j are needed to identify the tails of the distribution. Unlike methods

¹⁹There is some work on multinomial discrete choice models examining the nonparametric identification of the distribution of a choice-specific error $\epsilon_{\theta,j}$ and related parameters in models without random coefficients, where $\theta = \bar{\theta}$ for all types (Manski, 1975; Thompson, 1989; Matzkin, 1993; Lee, 1995). There is a larger literature on the binary choice and ordered choices models, such as Manski (1975), Cosslett (1983) and many others.

²⁰To avoid confusing literature references, the general linear independence strategy was introduced into statistics in a slightly later paper, Teicher (1963).

²¹The first version of our paper used reducibility to show identification of the distribution of heterogeneity for the multinomial choice model with subutility for choice j of $x_j \beta_\theta^j + w_j$. We then discovered that the same proof techniques could be applied, with almost no change, to the class of subutility functions $u_\theta^j(x_j) + w_j$, where θ indexes a member of the class of real analytic functions. Of course, multinomial choice is just one application of our general strategy for showing identification in nonlinear economic models.

such as Lewbel (2000), in both the simultaneous contributions of Berry and Haile and our paper, high covariate values are only needed to identify the tails of the distribution.

Studying the special case of $J = 2$, Ichimura and Thompson (1998) use the Cramer and Wold (1936) theorem for identification, which relies critically on a linear index functional form: $u_{\theta,j} = x'_j \beta_{\theta}$. We use only the quasilinearity of $u_{\theta}^j(x) + w_j$ in w_j and the real analytic assumption on $u_{\theta}(x)$. The key assumption is monotonicity in w_j . Ichimura and Thompson also need full support on all covariates to apply the Cramer-Wold theorem. Further, Ichimura and Thompson need an identification condition that reduces to our monotonicity condition that the sign of w_j in $u_{\theta}^j(x) + w_j$ is known. We need large support on only w . Gautier and Kitamura (2007) provide a computationally simpler estimator for the model of Ichimura and Thompson.

6 Selection Models

Selection models are one of the key tools in empirical microeconomics. More recent versions of these models, such as Heckman and Honore (1990) and Heckman (1990), emphasize that agents may sort on the heterogeneous returns to adopting some innovation. In our framework, this heterogeneity will arise as random coefficients or random functions in the outcome. We will generalize the earlier selection research and allow heterogeneity in both the selection and outcome equations.

6.1 Selection Models as Mixtures

Let there be $J \geq 2$ exclusive, discrete outcomes. Say these are competing products in a market. If a consumer chooses product $j \in J$, then we observe y^j , the quantity of product j purchased by the consumer. However, we do not observe y^k , the quantity of product k the consumer would have purchased if the consumer had picked product k . Thus, the quantity choice is a selected outcome as the researcher observes data on y_j only when product j is picked.

To be more concrete, suppose the potential outcomes of a type $\theta \in \Theta$ agent are such that $y_{\theta}^j = f_{\theta}^j(x_j)$, for $j = 1, \dots, J$ and $x_j \in X \subset \mathbb{R}^K$, where X contains an open and connected set. As before, we simply take X to be open and connected. Furthermore let $x = \{x_1, \dots, x_J\} \in X^J$. We will assume that an agent's vector of potential outcomes $f_{\theta} : X^J \rightarrow \mathbb{R}^J$ is a vector of real analytic functions. In a labor context, $f_{\theta}^j(x_j)$ can be interpreted as the function giving the economic return that an agent θ earns from choosing sector j . Another natural interpretation is that $f_{\theta}^j(x_j)$ represents the selling income a seller would earn if he chose selling mechanism j . This seller could be an auctioneer choosing among auction formats, in which case $f_{\theta}^j(x_j)$ would be the expected revenue he would earn in format j . One could study the seller of a house, and the possible selling mechanisms could be to sell through a real-estate agent, which would yield a selling price of $f_{\theta}^{j1}(x_1)$, or to sell without an agent, which would yield a return of $f_{\theta}^2(x_2)$.

As in the standard selection problem, we can only observe each type θ 's outcome y_θ^j for that sector j that θ chooses. We model this choice of sector through a selection equation, which is simply a multinomial choice over sectors. In particular, the utility to each mechanism j could depend upon a vector of covariates $(z_j, w_j) \in \mathbb{R}^{L+1}$. Following our analysis of identification in multinomial choice, we use a quasilinear representation of preferences over mechanisms, $u_\theta^j = v_\theta^j(z_j) + w_j$ where v_θ^j is a real analytic function. We make no assumption about “instruments”: x_j can be the same as $\{z_j, w_j\}$, or economic theory can be used to suggest that certain variables should enter the outcome equation vector, x_j , that are excluded from the discrete choice. Other variables, like tuition in a schooling choice context, might enter the discrete choice function but not the real-valued outcome.

To show reducibility, something about the continuous outcome must affect the discrete choice. Otherwise, if two types θ_1 and θ_2 are currently picking choice j and differ only in their real-valued outcome in sector k , we cannot prove that there is an experiment in the data where one (and only one) of the types switches to choice k . Therefore, we require a signal assumption: if $g_{\theta_1}^j(\cdot) \neq g_{\theta_2}^j(\cdot)$ for at least one choice j , then $v_{\theta_1}^k(\cdot) \neq v_{\theta_2}^k(\cdot)$ for at least one choice k . This is a restriction on the joint distribution of types, $G(\theta)$. The signal assumption may seem arbitrary, but there are important special cases that show its economic relevance. One special case is when the real-valued outcome itself enters the discrete choice, as in $v_\theta^j(z_j) = \beta_\theta g_\theta^j(x_j) + \tilde{v}_\theta^j(z_j)$, with $\beta_\theta \neq 0$. Alternatively, we could have $v_\theta^j(z_j) = s_\theta^j + \tilde{v}_\theta^j(z_j)$, where s_θ^j is an unobserved signal and where $g_{\theta_1}^j(\cdot) \neq g_{\theta_2}^j(\cdot)$ implies $s_{\theta_1}^j \neq s_{\theta_2}^j$.

To close the model, we assume the tie-breaking rule that type θ picks choice $j \in J$ only if θ strictly prefers j to all choices $j + 1, \dots, J$. As in the pure multinomial choice model, we assume there is an outside good with discrete-choice payoff components $v_\theta^0(\cdot) = 0$ and $w_0 = 0$. Our goal is to identify the joint distribution of $G(\theta)$ of types in the population. A type θ indexes the outcome functions $(g_\theta^1(x_1), \dots, g_\theta^J(x_J))$ and the subutility functions $v_\theta^j(z_j)$. It is essential to allow the random outcome functions $g_\theta^j(x_1)$ to be correlated with the random functions in the selection equation, $v_\theta^j(z_j)$.

Suppose the researcher has data on

$$H(j, y, z, w) = \Pr(D = j, Y \leq y \mid (z, w))$$

for all combinations of j and y and covariates $(z, w) \in Z \times W$. Our approach is to model an agent's selection decision and outcome realization as a vector-valued response. Given the covariates (z, w) of the model, $\Pr(D = j, Y \leq y \mid (z, w))$ is the expectation of

$$1[D = j, Y \leq y] = 1\left[v_\theta^j(z_j) + w_j \geq v_\theta^k(z_k) + w_k \forall k \neq j\right] \cdot 1\left[g_\theta^j(x_j) \leq y\right] \quad (8)$$

with respect to the distribution $G(\theta)$. Following the definition of I -sets for such models described in Section 3, for any finite set of types $T = \{\theta_t\}_{t=1}^N \subset \Theta$ and any combination of (j, y, z, w) , let

$I_{j,y,v}^T$ index the subset of types for whom the event (8) is satisfied,

$$I_{j,y,z,w,x}^T = \left\{ \theta \in T \mid \left[v_{\theta}^j(z_j) + w_j \geq v_{\theta}^k(z_k) + w_k \forall k \neq j \right] \& \left[g_{\theta}^j(x_j) \leq y \right] \right\}. \quad (9)$$

Theorem 9. *The distribution G over types $\theta \in \Theta$ is identified within $\mathcal{G}$ in the selection model.*

Proof. Consider a non-empty

$$I_{0,\infty,z,w,x}^T = \left\{ \theta \in T \mid 0 > v_{\theta}^j(z_j) + w_j \forall j \neq 0 \right\},$$

which we can always assure by setting each w_j sufficiently low to induce some type in T to pick the outside good. If $I_{0,\infty,z,w,x}^T$ has two or more elements, let $\theta_1, \theta_2 \in I_{0,\infty,z,w,x}^T$ and $\theta_3 \in T \setminus I_{0,\infty,z,w,x}^T$.

The first case is when $v_{\theta_1}^j(z_j) > v_{\theta_2}^j(z_j)$ for some choice j . Then we can increase w_j to w'_j to induce $\theta_1 \in T \setminus I_{0,\infty,z,w',x}^T$ but $\theta_2 \in I_{0,\infty,z,w',x}^T$, where w' is w with w'_j replacing w_j in the j th slot of w . By monotonicity, $\theta_3 \in T \setminus I_{0,\infty,z,w',x}^T$.

The second case is when $v_{\theta_1}^j(z_j) > v_{\theta_2}^j(z_j)$ at z but the functions differ. Then there is an open set U around z_j where θ_1 and θ_2 pick 0. There is a point z'_j and $v_{\theta_1}^j(z_j) \neq v_{\theta_2}^j(z_j)$, by the properties of real analytic functions. If necessary to keep θ_3 from picking choice 0, w 's for other choices can be increased. Then we are back in the first case, with the new z having z'_j for the j th choice.

The third case could be when the functions $v_{\theta_1}^j(\cdot) = v_{\theta_2}^j(\cdot)$ for all choices $j = 1, \dots, J$, but differ in at least one of the sector-specific outcome functions, $g_{\theta}^j(\cdot)$. We rule this case out by the signal assumption: if $g_{\theta_1}^j(\cdot) \neq g_{\theta_2}^j(\cdot)$, then $v_{\theta_1}^k(\cdot) \neq v_{\theta_2}^k(\cdot)$ for some choice k , which returns us to the first and second cases. $\square$

The proof is really no different than the proof of Theorems 7 and 8, for the pure discrete choice model. This is because of the signal assumption: if types differ in their sector-specific outcome equations, they must differ in their discrete choice utilities. As stated above, this assumption nests the case where discrete choice utilities include the real-valued payoffs, such as workers considering their wages before making a sector choice. If we did not include the signal assumption, then we could not show reducibility: two types in an I -set $I_{0,\infty,z,w,x}^T$ could have identical discrete choice payoffs but different real-valued functions $g_{\theta_1}^j(\cdot)$ for some other choice j . Then both types would switch to choice j at the same point, so we could not keep one type in $I_{0,\infty,z,w,x}^T$ while the other leaves. This shows reducibility may be stronger than is necessary for identification. However, we are sure extending our proof to the case where sector choice is completely unrelated to outcomes in each sector is economically interesting.

6.2 Roy Model: Covariates Only in The Outcome Equation

Now we will study the Roy model, where an agent simply chooses the sector with the highest real-valued response. Let w_j be a scalar covariate that enters outcome j : $y_\theta^j = f_\theta^j(x_j) + \beta_\theta^j w_j$ for $\{x_1, \dots, x_J\} \in X \subset \mathbb{R}^{JK}$, where X is open and connected. Let $w = \{w_1, \dots, w_J\} \in \mathbb{R}_+^J$. The key monotonicity assumption is that $\beta_\theta^j > 0$ for all $\theta \in \Theta$. Furthermore let selection be the Roy model: agent θ picks choice j when $y_\theta^j \geq y_\theta^k \forall k \neq j$. Thus a type θ indexes a pair $\left\{ \left\{ f_\theta^j(\cdot) \right\}_{j=1}^J, \left\{ \beta_\theta^j \right\}_{j=1}^J \right\}$, where we impose the restriction as before that each $f_\theta^j : X \rightarrow \mathbb{R}$ is real analytic. Note that we do not allow for an outside good, although an outside good (unemployment, say) could easily be added. We use the tie-breaking rule that choice $j \in J$ is picked only if it is strictly preferred to all choices $j+1, \dots, J$.

Theorem 10. *The joint distribution G of θ is identified in $\mathcal{G}$ for the Roy model.*

Proof. We appeal to Theorem 3, our theorem on two step identification. We know from the continuous-response Theorem 4 that the submodel where all J sector-specific returns y_j^θ are observed is reducible, and hence identifiable. We only need to show reducibility for the submodel where each submodel-type $\tilde{\theta}$ is given by a vector of potential outcomes $(\alpha_{\tilde{\theta}}^1, \dots, \alpha_{\tilde{\theta}}^J) \in \mathbb{R}^J$ and a vector of coefficients $(\beta_{\tilde{\theta}}^1, \dots, \beta_{\tilde{\theta}}^J) \in \mathbb{R}_+^J$, where each type chooses product j if and only $\alpha_{\tilde{\theta}}^j + \beta_{\tilde{\theta}}^j w_j \geq \alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k, \forall k \neq j$, and where the covariates are allowed to vary as $w = (w_1, \dots, w_J) \in \mathbb{R}_+^J$. As is usual in selection, $y_{\tilde{\theta}}^j = \alpha_{\tilde{\theta}}^j + \beta_{\tilde{\theta}}^j w_j$ is only observed when sector j is picked.

We show reducibility of this second stage submodel. Let T be a finite set of types. Let $\{1, y, w\}$ be a triplet that defines

$$I_{1,y,w}^T = \left\{ \tilde{\theta} \in T \mid \alpha_{\tilde{\theta}}^1 + \beta_{\tilde{\theta}}^1 w_1 \leq y \text{ \& } \alpha_{\tilde{\theta}}^1 + \beta_{\tilde{\theta}}^1 w_1 > \alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k \text{ for } k = 2, \dots, J \right\}$$

with at least two elements. Let two of those elements be $\tilde{\theta}_1, \tilde{\theta}_2 \in I_{1,y,w}^T$. Let $\tilde{\theta}_3 \in T \setminus I_{1,y,w}^T$. There exists a non-empty $I_{1,y,w}^T$ because the w_1 and y can always be increased (and other w_k 's decreased).

The first case is when $y_{\tilde{\theta}_1}^1 > y_{\tilde{\theta}_2}^1$. Let $y' = y_{\tilde{\theta}_2}^1$, so that $\tilde{\theta}_2 \in I_{1,y',w}^T$, $\tilde{\theta}_1 \notin I_{1,y',w}^T$ and $\tilde{\theta}_3 \notin I_{1,y',w}^T$.

The second case is when $y_{\tilde{\theta}_1}^1 = y_{\tilde{\theta}_2}^1$ but $\{\alpha_{\tilde{\theta}_1}^1, \beta_{\tilde{\theta}_1}^1\} \neq \{\alpha_{\tilde{\theta}_2}^1, \beta_{\tilde{\theta}_2}^1\}$. Recall $\tilde{\theta}_1$ and $\tilde{\theta}_2$ strictly prefer choice 1. Reduce w_1 to w'_1 by a small amount, so that at least one of $\tilde{\theta}_1, \tilde{\theta}_2$ remains in $I_{j,y,w'}$, where w' is the vector w with w'_1 in the first slot. If both $\tilde{\theta}_1, \tilde{\theta}_2 \in I_{1,y,w'}$, then let $y_{\tilde{\theta}_1}^{1'} = \alpha_{\tilde{\theta}_1}^1 + \beta_{\tilde{\theta}_1}^1 w'_1$ and $y_{\tilde{\theta}_2}^{1'} = \alpha_{\tilde{\theta}_2}^1 + \beta_{\tilde{\theta}_2}^1 w'_1$. Because straight lines can cross at most once, $y_{\tilde{\theta}_1}^{1'} \neq y_{\tilde{\theta}_2}^{1'}$. Without loss of generality, say $y_{\tilde{\theta}_1}^{1'} < y_{\tilde{\theta}_2}^{1'}$. Set $y' = y_{\tilde{\theta}_1}^{1'}$. Then $\tilde{\theta}_1 \in I_{1,y',w'}$, $\tilde{\theta}_2 \notin I_{1,y',w'}$, and, by monotonicity, $\tilde{\theta}_3 \notin I_{1,y',w'}$.

The third case is when $\{\alpha_{\tilde{\theta}_1}^1, \beta_{\tilde{\theta}_1}^1\} = \{\alpha_{\tilde{\theta}_2}^1, \beta_{\tilde{\theta}_2}^1\}$ but $\{\alpha_{\tilde{\theta}_1}^k, \beta_{\tilde{\theta}_1}^k\} \neq \{\alpha_{\tilde{\theta}_2}^k, \beta_{\tilde{\theta}_2}^k\}$ for some other

type k . There are values $w_k^{\tilde{\theta}_1}$ and $w_k^{\tilde{\theta}_2}$ that solve

$$\alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k = \alpha_{\tilde{\theta}}^1 + \beta_{\tilde{\theta}}^1 w_1 \quad (10)$$

for w_k , for $\tilde{\theta}_1$ and $\tilde{\theta}_2$, respectively. Say $w_k^{\tilde{\theta}_1} > w_k^{\tilde{\theta}_2}$. Take $w'_k = w_k^{\tilde{\theta}_2}$ and w' to be w with the k th element replaced by w'_k . Then $\tilde{\theta}_1 \in I_{1,y',w'}$, $\tilde{\theta}_2 \notin I_{1,y',w'}$, and, by monotonicity, $\tilde{\theta}_3 \notin I_{1,y',w'}$. Now say $w_k^{\tilde{\theta}_1} = w_k^{\tilde{\theta}_2}$. Perturb w_1 to w'_1 so that at least one of $\theta_1, \theta_2 \in I_{j,y,w'}$. Because two straight lines, here $\alpha_{\tilde{\theta}_1}^k + \beta_{\tilde{\theta}_1}^k w_k$ and $\alpha_{\tilde{\theta}_2}^k + \beta_{\tilde{\theta}_2}^k w_k$, can intersect at most once, the solution to (10), with w'_1 replacing w_1 , cannot involve $w_k^{\tilde{\theta}_1} = w_k^{\tilde{\theta}_2}$. If $w_k^{\tilde{\theta}_1} > w_k^{\tilde{\theta}_2}$. Take $w''_k = w_k^{\tilde{\theta}_2}$, and $I_{j,y,w''}$ reduces $I_{j,y,w'}$, which satisfies $I_{j,y,w'} \subseteq I_{j,y,w}$. $\square$

6.3 Selection Literature Review

We model the joint decision of what product to choose and the usage conditional on that product as a vector-valued outcome. Therefore, we place our identification problem in the mathematical structure of a mixtures model: (2). We use reducibility and hence linear independence to show identification of both the continuous outcome (usage) and discrete selection (product choice) equations.

To some degree, our approach to selection hearkens back to the original literature on selection in the 1970s (Gronau, 1974; Heckman, 1974, 1979). This literature made parametric assumptions about the error terms in the model, and then suggested maximum likelihood as the estimation method. More recent selection papers moved away from this full identification approach to focus on the relationship between selection models and instrumental variables, among other issues. By contrast, we return to the full identification approach and identify, nonparametrically, the joint distributions of all random components. The dependent variable we model is the joint outcome, (j, y) : the discrete choice j and the continuous choice y . We impose the full structure of the selection model are able to show identification for the complete model.

6.3.1 Full Identification, Treatment Effects and Other Calculable Economic Measures

We use our reducibility framework to prove the full nonparametric identification of random coefficients in all parts of the selection model. In particular, we identify the full joint distribution of all intercepts and slopes in both the outcome and selection equations. Identifying the full distribution of outcomes in either the Roy model or in the more standard selection model allows us to compute any treatment effect possible. Selection models sometimes fail to identify the full joint distribution of the unobservables in the model. For example, if e_1 is the unobserved error (“treatment effect”) in the continuous outcome for choice 1 and e_2 is the unobserved error in the continuous outcome

for choice 2, the identification at infinity framework in Heckman (1990) does not allow the identification of the joint density of e_1 and e_2 . Heckman, Smith and Clements (1997) explore many methods that might be able to identify the full distribution of outcomes, but are eventually unable to do so. Heckman and Honore (1990) do identify the joint distribution for the Roy model, which is by definition a selection model without covariates in the selection equations that are excluded from the outcome equations. In our model, the full distribution G of all unobservables, in both the selection and outcome equations and including outcome intercepts (the treatment effects), is identified.

Even though they do not prove full identification, Heckman, Smith and Clements (1997) is a good reference for the importance of being able to compute any desired function of the outcomes, not just the mean treatment effects $E[y_1 - y_2 | x]$ often focused on in the literature. One immediate idea is to focus on medians and not means, as in median $[y_1 - y_2 | x]$, as medians may be less sensitive to the tail frequencies of y_1 and y_2 . The traditional literature does not focus on medians, while our paper fully identifies the model, so the median treatment effect is easily identified.

Consider a firm adopting a new technology. If y_1 is the outcome for adopting, picking choice 1, and y_2 is the outcome for not adopting, picking choice 2, then some researchers might be interested in the fraction of firms that would benefit from the technology at zero cost, or $\Pr(y_1 < y_2 | x)$. This is a feature of the full joint distribution G of all the random parameters in the model. Another idea to consider is whether a high $E[y_1 - y_2 | x]$ is due to a high $E[y_1 | x]$ or a low $E[y_2 | x]$. Again, the approach of fully identifying the model allows us to compute these measures.

6.3.2 Allowing Random Coefficients in the Selection Equation

Following Heckman (1990) and Heckman and Honore (1990), a large literature has focused on the assumptions on both the data and model allowing nonparametric identification of the distribution of heterogeneity in an outcome equation: those with a high return to college are more likely to attend college.²²

Recently, Heckman, Urzua and Vytlačil (2006) have mentioned that selection models allow random coefficients (heterogeneity) only in the outcome equation, not the selection equation. “Essential heterogeneity” is allowed for only part of the model. In industrial organization demand estimation, the product selection equation is often a random coefficients logit when quantity choice is not modeled. The elasticities implied by the alternative homogeneous coefficients logit model are restrictive. Adding random coefficients to the selection equation allows consumers to substitute to observably similar products, even if quantity choice is also in the model. Unlike previous identification results for selection models, our results allow random coefficients in both the selection and

²²See for example, Imbens and Angrist (1994), Heckman and Vytlačil (1999), Manski and Pepper (2000), Lewbel (2000), Sørensen (2006), Heckman and Navarro (2007), Florens, Heckman, Meghir and Vytlačil (2008), Wooldridge (2007) and many others.

outcome equations.

Allowing random coefficients in all aspects of the model has other benefits. For example, a covariate in the selection equation may induce some agents to substitute towards choice 1 and for others to substitute away from choice 1. There is no reason the signs in the selection equation should be the same for all types. In the identification theorems, we usually require one variable w , in either the outcome or selection equations, to have support that is either strictly negative or strictly positive. Even with a common signs, allowing for random coefficients on this variable means the magnitude of the effect of substituting towards or away from various options can vary quite a bit across the population.

6.3.3 Multinomial Choice

Many selection papers rely on a structure with only two groups, say college or non-college or treatment and control. To some extent, methods that rely on identification at infinity ask even more of the data to be able to move the probability of one out of J exclusive outcomes to be 1. By contrast, our mixtures identification approach relies on linear independence and so selection generalizes easily to the case of multinomial choice in the selection equation. This is important for applications to demand estimation in industrial organization, where often the selection decision is the choice between more than two brands, including the option of no purchase. For example, in an environmental application, a household could pick between a central air conditioner, one or more room air conditioners and the outside option of no air conditioner before making their continuous choice of what temperature to cool a house to.

6.3.4 Why No Identification at Infinity?

Recently, Heckman, Urzua and Vytlačil (2006) have discussed that identification in other selection frameworks relies on finding data on x where the probability of selection, say part of $H(y, x)$, is 1 or 0. This is called “identification at infinity”, and is thought to be problematic in finite samples as this type of data is often not available. Chamberlain (1986), Andrews and Schafgans (1998) and Khan and Tamer (2007) show that many estimators of a finite vector of parameters based on identification at infinity arguments have slower than $\sqrt{n}$ rates of convergence if the support of the error terms is not smaller than the support of the exogenous data and higher moments are finite. Intuitively, identification at infinity relies on small slices of data that move the selection probabilities towards 1 or 0, so the estimator converges more slowly than estimators that make full use of the data. Our view of selection models as a vector-valued outcome removes any need to consider identification at infinity. We model how types respond differently at different x ’s, rather than trying to find some x where all the types have the same response, like all choosing the same

sector for employment.²³

Our proofs of the identification of the selection model use reducibility and hence linear independence, not identification at infinity. Identification at infinity requires that the probability of some choice j be set to 1, not 0. This involves properties of the discrete component j of the vector of outcomes $\{y_j, j\}$. As the proof of Theorem 2 shows, given an arbitrary set of types T with n elements, we need to be able to reduce the set of types purchasing a good to a singleton. It does not really matter what the other $n - 1$ types in the arbitrary, finite set T are doing. In a multinomial choice setting, this is far from setting the probability of some choice j to 1. This means all the types are taking the same action. Further, in identification at infinity, the estimator mainly uses the small slices of data where the probability of a choice is close to 1. Using linear independence, all the data is used, because types enter or leave I -sets all of the time.

Let us try to explain our understanding of the term identification at infinity in more detail. For expositional purposes, consider a simple selection model where

$$1[D = j, Y \leq y] = 1[z'_j \beta \geq z'_k \beta \forall k \neq j] \cdot 1[x'_j \gamma \leq y],$$

where z_j are the only parameters in the selection model and x_j are the only parameters in the outcome equation. In a typical two-step approach to identification in selection models, researchers view the selection equation covariates $\{z_j\}_{j=1}^J$ as a set of instruments. The researcher looks at the conditional mean of $y_j = x'_j \gamma$ given selection, or

$$E[y_j | D = j, \{z_j, x_j\}_{j=1}^J] = x'_j E[\gamma | D = j, \{x_j, z_j\}_{j=1}^J].$$

The researcher then algebraically derives a formula for $E[\gamma | D = j, \{x_j, z_j\}_{j=1}^J]$ using the structure of the model. This algebraic formulation gives a correction term that allows the consistent estimation of the unknown parameters.

In a typical simple application, γ might be a set of homogeneous coefficients, except for a random intercept ϵ_j . Denote the non-random parameters with tildes. In this case,

$$E[y_j | D = j, \{x_j, z_j\}_{j=1}^J] = \tilde{\gamma}_{0,j} + x'_j \tilde{\gamma} + E[\epsilon_j | D = j, \{x_j, z_j\}_{j=1}^J],$$

where here $\tilde{\gamma}_{0,j}$ is the true, non-random intercept for product j and ϵ_j is an error term with an unconditional mean of 0. One identification problem is that it is difficult to distinguish the true intercept $\tilde{\gamma}_{0,j}$ and the expected value of the error term in the selected sample $E[\epsilon_j | D = j, \{x_j, z_j\}_{j=1}^J]$. However, if the researcher can find a value of $\{z_j\}_{j \in J}$ so that $\Pr(D = j) = 1$, then usually

²³For special cases of selection and endogeneity, Hong and Tamer (2003), Vytlacil and Yildiz (2007), and Shaikh and Vytlacil (2005) do not rely on identification at infinity. However, these papers do not identify the full distribution of unobserved heterogeneity and do not allow essential heterogeneity in the selection equation.

$E[\epsilon_j \mid D = j, \{x_j, z_j\}_{j=1}^J] = 0$ as all consumers choose product j and there is no selection issue. Then $\tilde{\gamma}_{0,j}$ can be identified separately from the conditional mean of ϵ_j for this value of $\{z_j\}_{j \in J}$.

Our mixtures approach to this problem avoids the need to find characteristics $\{x_j, z_j\}_{j=1}^J$ that set $\Pr(D = j) = 1$. In the above example with only a random choice-specific intercept, we would be interested in identifying the distribution G of $(\{\epsilon_j\}_{j=1}^J)$, the J intercepts and the random coefficients in the selection equation, in addition to the homogeneous parameters $\tilde{\gamma}$ and $\{\tilde{\gamma}_{0,j}\}_{j \in J}$. We simply need to show that a unique distribution G and set of homogeneous parameters solves the mixture integral equation.

Heckman and Vytlačil (2001) present bounds for the average treatment effect for the case of $J = 2$

$$E[y_2 \mid \{x_j, z_j\}_{j=1}^J] - E[y_1 \mid \{x_j, z_j\}_{j=1}^J].$$

There is no conditioning on the endogenous outcome D in the definition of the average treatment effect. Their bounds point identify the average treatment effect under identification at infinity. If the exogenous data does not allow the researcher to observe a situation where $\Pr(D = 2) = 1$, Heckman and Vytlačil prove their bounds are sharp: a model can be constructed that is 1) consistent with the data on $\{Y, D, \{x_j, z_j\}_{j=1}^2\}$ and 2) generates any particular value of the average treatment effect inside the bounds. One interpretation of the sharpness theorem is that the average treatment effect may not be point identified under their assumptions. While Heckman and Vytlačil do not allow random coefficients in the selection equation, they are nonparametric about how the observed covariates $\{z_j\}_{j=1}^2$ enter the selection equation. Their selection rule is

$$D = 2 \iff \mu\left(\{z_j\}_{j=1}^2\right) \geq 0.$$

This model has no random coefficients or essential heterogeneity in the selection equation. We allow random coefficients, and indeed allow a nonparametric distribution G for those random coefficients, but we impose some structure in the selection subutility functions. We believe our limited structure on the selection equation allows us to gain identification from working with the joint probability of y and D . Our assumptions and those of Heckman and Vytlačil are non-nested.

7 Estimation: A Fake Data Experiment for Selection

This paper is about identification, not estimation. Because we have established identification, any of the mixtures estimators listed in the introduction could (up to regularity conditions) be used for consistent estimation of G , the distribution of random coefficients. However, our linear independence identification strategy naturally suggests the linear regression mixtures estimator of Bajari, Fox, Kim and Ryan (2007). The fake data experiment is a little beyond our current

identification theorems, but does demonstrate the power of mixtures in the selection model. The selection literature has focused a lot on treatment effects and how the requirements of identification at infinity makes identifying treatment effects hard in a finite sample. The main goal is to see whether the distribution of the intercepts of two brands can be recovered nonparametrically. Also, we extend the selection literature and allow a random coefficient in the selection equation.

In the first stage, an agent i chooses product $j = 1, 2$ if $u_{i,j} \geq u_{i,k}$, $k \neq j$ and where $u_{i,j} = x_{i,j,1}\beta_1$. In the second stage, the choice-specific continuous outcome satisfies $y_{i,j} = x_{i,j,2} + \beta_{j+1}$, and is only observed for the chosen j in the first stage. There are three random coefficients, a slope in the selection equation, β_1 , and two product-specific constants in the outcome equations, β_2 and β_3 . The random coefficients have the asymmetric mixed normal distribution

$$g(\beta_1, \beta_2, \beta_3) = 0.4 \cdot N\left(\begin{bmatrix} -3 \\ -2 \\ -3 \end{bmatrix}, \begin{bmatrix} 3 & 2 & -0.6 \\ 2 & 3 & 0.3 \\ -0.6 & 0.3 & 2.1 \end{bmatrix}\right) + 0.6 \cdot N\left(\begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 & 1.5 & 0.2 \\ 1.5 & 2 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}\right).$$

Our goal is to estimate $g(\beta_1, \beta_2, \beta_3)$. Using identification at infinity strategies, it is usually thought to be difficult to estimate the joint distribution of the intercepts in a selection model (Heckman, 1990). There are four exogenous covariates with distribution

$$f\left(\begin{bmatrix} x_{1,1} \\ x_{2,1} \\ x_{1,2} \\ x_{2,2} \end{bmatrix}\right) = N\left(\begin{bmatrix} 1 \\ 1 \\ 2 \\ -1.3 \end{bmatrix}, \begin{bmatrix} 3 & 0 & 1.6 & -1 \\ 0 & 3 & 1.6 & -1 \\ 1.6 & 1.6 & 5 & 0.1 \\ -1 & -1 & 0.1 & 2 \end{bmatrix}\right).$$

The researcher lacks data on $x_{2,2}$ if $j = 1$ is chosen, and so forth. Indeed, in our fake data example there are $i = 1, \dots, N = 15,000$ observations on $(j_i, y_{i,j_i}, x_{i,1,1}, x_{i,2,1}, x_{i,j_i,2})$ for individuals i .

There is selection bias in this example. The true mean of β_2 , the intercept for choice 1, is $0.4 \cdot (-2) + 0.6 \cdot 3 = 1$, but an OLS regression of $y_{i,1}$ on $x_{1,2}$ for the sample with $j_i = 1$ gives 0.54 for the intercept. Therefore, regression understates the mean or average treatment effect of 1 by 0.46 points. Because of the large number of observations, this outcome is unlikely to be due to sampling error. Likewise, the true mean of β_3 is $0.4 \cdot (-3) + 0.6 \cdot 2 = 0$, while OLS of y_{i2} on $x_{2,2}$ gives -0.50, understating the mean by 0.50. Likewise, the pooled regression is also inconsistent for the pooled mean intercept. The true pooled mean intercept is $0.5 \cdot 1 + 0.5 \cdot 0 = 0.5$, as our i.i.d. sampling scheme makes choosing product 1 and 2 equally likely. The pooled OLS regression of the observed y_i on the observed $x_{-,2}$ gives 0.02 for the intercept, understating the mean treatment effect by 0.48.

Our estimator for $f(\beta_1, \beta_2, \beta_3)$ is the nonparametric mixtures estimator of Bajari et al.. The mixtures estimator uses a discrete grid of $R = 75$ vectors $\beta^r = (\beta_1^r, \beta_2^r, \beta_3^r)$. Each grid point β^r is

the mean of a normal distribution centered around that grid point. Given an estimated weight $\hat{\theta}^r$ on each of the R normal components, the final estimator for $g(\beta_1, \beta_2, \beta_3)$ is

$$\hat{g}(\beta_1, \beta_2, \beta_3) = \sum_{r=1}^R \hat{\theta}^r \phi(\beta_1 - \beta_1^r) \phi(\beta_2 - \beta_2^r) \phi(\beta_3 - \beta_3^r),$$

where ϕ is the standard normal density. We use a Weyl sequence to pick the R means of the normals. Note that the individual normals are independent, while our target density $g(\beta_1, \beta_2, \beta_3)$ is a mixture of correlated multivariate normals.

Next we pick $s = 1, \dots, S$ points of evaluation (j^s, y_j^s) for the dependent variable. The estimator is implemented using the regression equation, for an arbitrary pair (j^s, y_j^s) and observation i

$$1[j_i^* = j^s, y_{i,j} \leq y_j^s] = \sum_{r=1}^R \theta^r \cdot \int_{\beta_1} \int_{\beta_2} \int_{\beta_3} H(\beta, x_i, j^s, y_j^s) \phi(\beta_1 - \beta_1^r) \phi(\beta_2 - \beta_2^r) \phi(\beta_3 - \beta_3^r) d\beta_1 d\beta_2 d\beta_3 \quad (11)$$

where the expression

$$H(\beta, x_i, j^s, y_j^s) = 1[x_{i,j^s,1}\beta_1 > x_{i,k,1}\beta_1 \forall k \neq j^s] 1[x_{2,i,j^s} + \beta_{1+j^s} \leq y_j^s]$$

represents whether an agent with preferences β would both select choice j and have a continuous outcome $y_j \leq y_j^s$, with the covariates of the statistical observation i . We choose $S = 2$, which corresponds to the 2 binary choices times 31 evenly spaced y_j^s points, from -10.5 to 12.7. The three-dimensional numerical integral in (11) only needs to be computed once; we use Gauss-Hermite quadrature because the density we are integrating against is an independent normal in each dimension.

If there are N statistical observations, there are $N \cdot R \cdot S$ regression observations in a linear regression where the only unknown parameters are the R θ^r 's. As the unknown parameters θ^r enter (11) linearly, the estimator is linear regression subject to the inequality constraints $\theta^r \geq 0 \forall r = 1, \dots, R$ and $\sum_{r=1}^R \theta^r = 1$. Linearly constrained linear regression is a convex optimization problem, where commonly available solvers are guaranteed to find a global minimum.

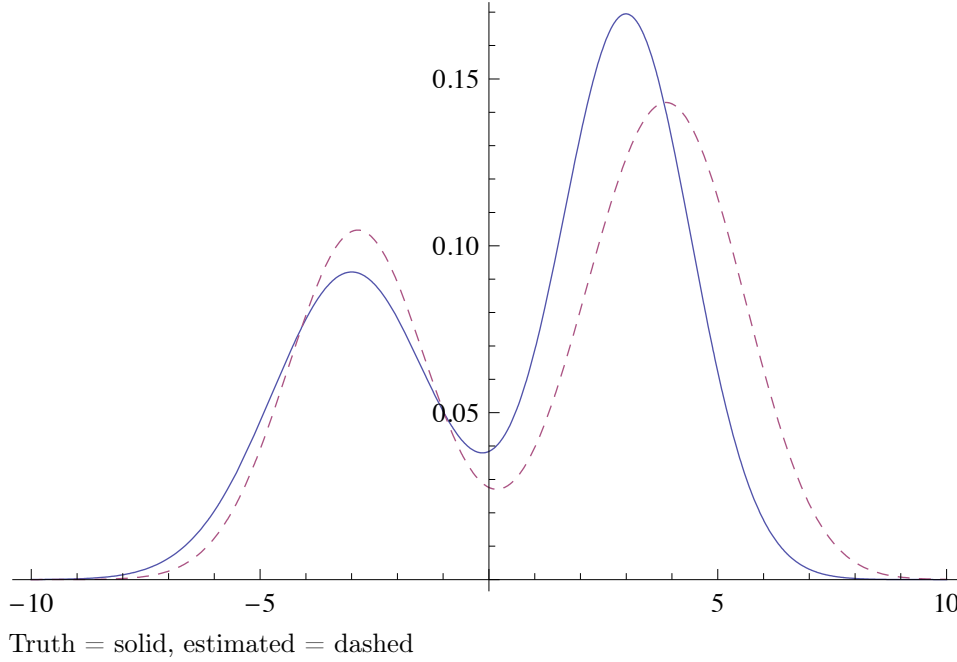
Table 1 reports the means of each of the three random coefficients, both from the truth, our earlier reported OLS results, and from our selection estimator. The table shows we get the means of the two continuous outcomes (treatment effects) about right: for outcome 1 the difference is 0.01 and for outcome 2 it is 0.04. Most importantly, our estimates of the mean treatment effects are far closer to the truth than the uncorrected OLS regressions. For the discrete choice slope coefficient, the estimate is off by more, but as we will see below, we get the shape of the marginal density of β_1 about right.

A measure of statistical fit of a density estimate (against the truth) is the root integrated

Table 1: True and Estimated Means of the Random Coefficients

	Truth	OLS	Selection method
Discrete choice β_1	0.6	N/A	1.16
Outcome 1 β_2	1	0.54	0.99
Outcome 2 β_3	0	-0.50	0.04

Figure 1: True and Estimated Marginal Density of the Slope in the Selection Equation, $\beta_{i,1}$



squared error. Here, the root integrated squared error is

$$\left(\left(\int \hat{g}(\beta_1, \beta_2, \beta_3) - g(\beta_1, \beta_2, \beta_3) \right)^2 d\beta_1 d\beta_2 d\beta_3 \right)^{1/2} = 0.07,$$

meaning that, across the support of the density, the true and estimated joint densities are off by a mean of 0.07. This strikes us as relatively low.

Figure 1 shows that the marginal distribution of the selection equation's slope, $\beta_{i,1}$, is well identified nonparametrically. The solid line is the truth and the dashed line is the estimate. Aside from the shift right (explaining the mean difference in Table 1), the shape of the marginal density is recovered nicely. To our knowledge, Figure 1 is the first instance of estimating random coefficients in the selection equation.

Figure 2: True and Estimated Marginal Density of the Intercept in the Outcome Equation For Choice 1, $\beta_{i,2}$

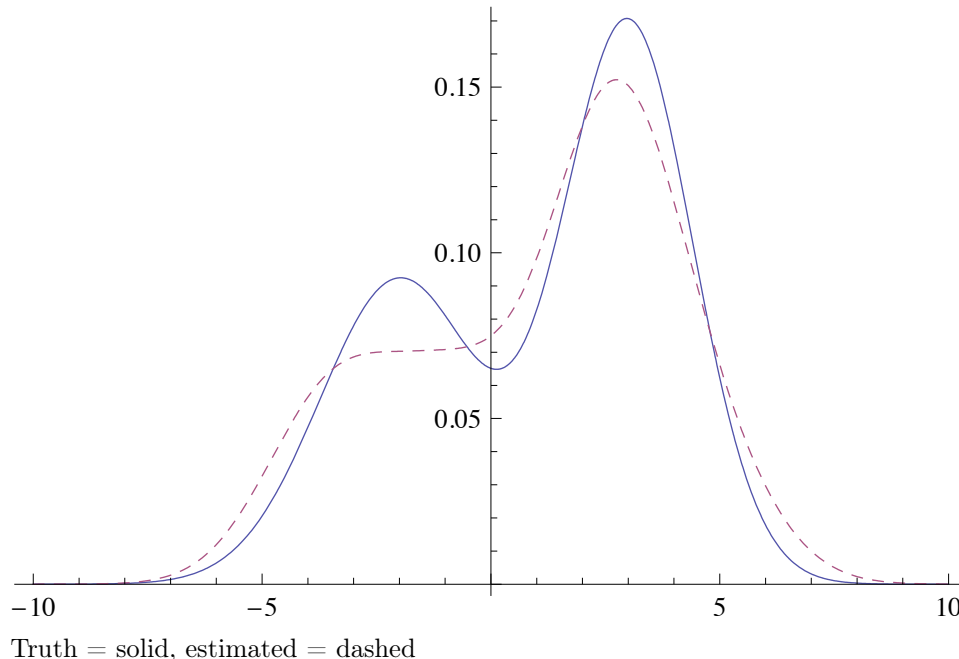


Figure 2 shows the truth (solid) and the nonparametrically estimated (dashed) marginal densities of the intercept in the outcome equation for choice 1. Again, our estimator nails the density even though identification of the intercept's distribution is thought to be difficult.

Figure 3 plots the true and estimated marginal densities of the intercept in the 2nd equation, β_3 . We get the modes correct, but understate the height of both modes. Still, this is pretty good.

Our estimator $\hat{f}(\beta_1, \beta_2, \beta_3)$ is a multivariate estimator. We cannot plot a density of three random variables, so we look at the estimated bivariate densities. Each figure plots both the truth and the estimated joint densities. The same scale is used for both plots in order to aid visual inspection. Figure 4 plots the joint density of the selection slope, β_1 , and the intercept for outcome 1, β_2 . The third box is the error, or $f(\beta_1, \beta_2, \beta_3) - \hat{f}(\beta_1, \beta_2, \beta_3)$. We see that the error lies between -0.02 and 0.02, which is generally good. For space reasons, we omit the plot of the bivariate density for β_1 and β_3 .

Heckman (1990) shows that an identification at infinity strategy cannot recover the joint distribution of the outcomes, $\{\beta_2, \beta_3\}$, because the researcher only observes data on only one of these outcomes at a time. Here we examine whether our linear independence identification strategy can identify the bivariate distribution of outcomes. Figure 5 plots the joint density of the two outcome

Figure 3: True and Estimated Marginal Density of the Intercept in the Outcome Equation For Choice 2, $\beta_{i,2}$

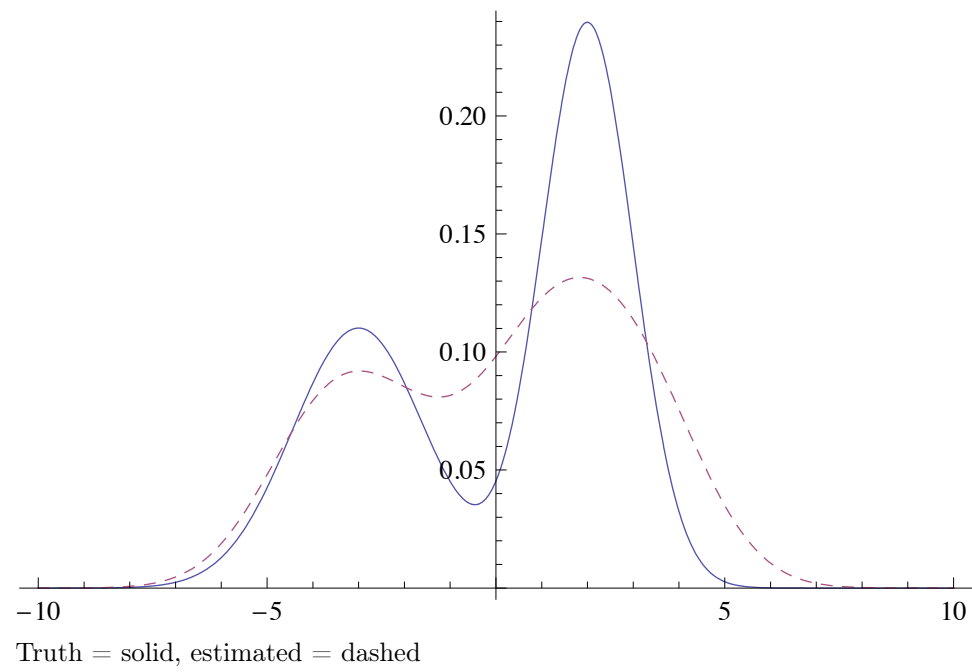


Figure 4: True and Estimated Joint Density of the Selection Slope and Intercept in the Outcome Equation For Choice 1, $\beta_{i,1}$ and $\beta_{i,2}$

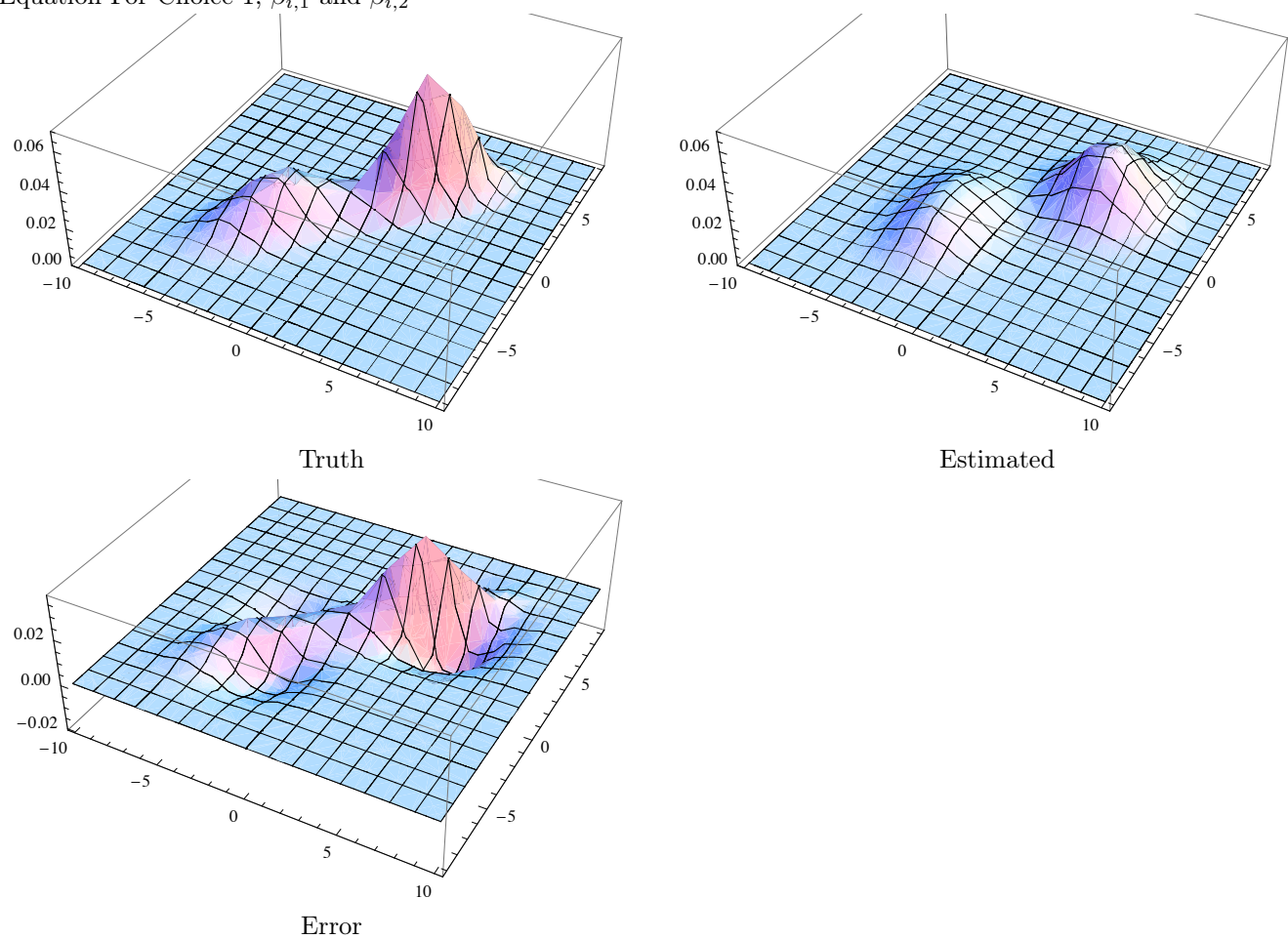


Table 2: True and Estimated Correlation Matrices of the Random Coefficients

Truth			Estimated		
1	0.92	0.75	1	0.73	0.74
0.92	1	0.79	0.73	1	0.65
0.75	0.79	1	0.74	0.65	1

These are the correlation matrices for, in order, the random slope coefficient in the discrete choice subutility, the random intercept for outcome 1, and the random intercept for outcome 2.

equations. It seems like we estimate the joint density well; again the maximum error is 0.02 density points.

Another way of addressing the success of the estimates of the three bivariate densities is to look at the implied correlation matrices. Table 2 presents the true and estimate correlation matrices. We see that, while not extremely accurate, the estimated correlations are in the correct ballpark of the true correlations.

8 Conclusions

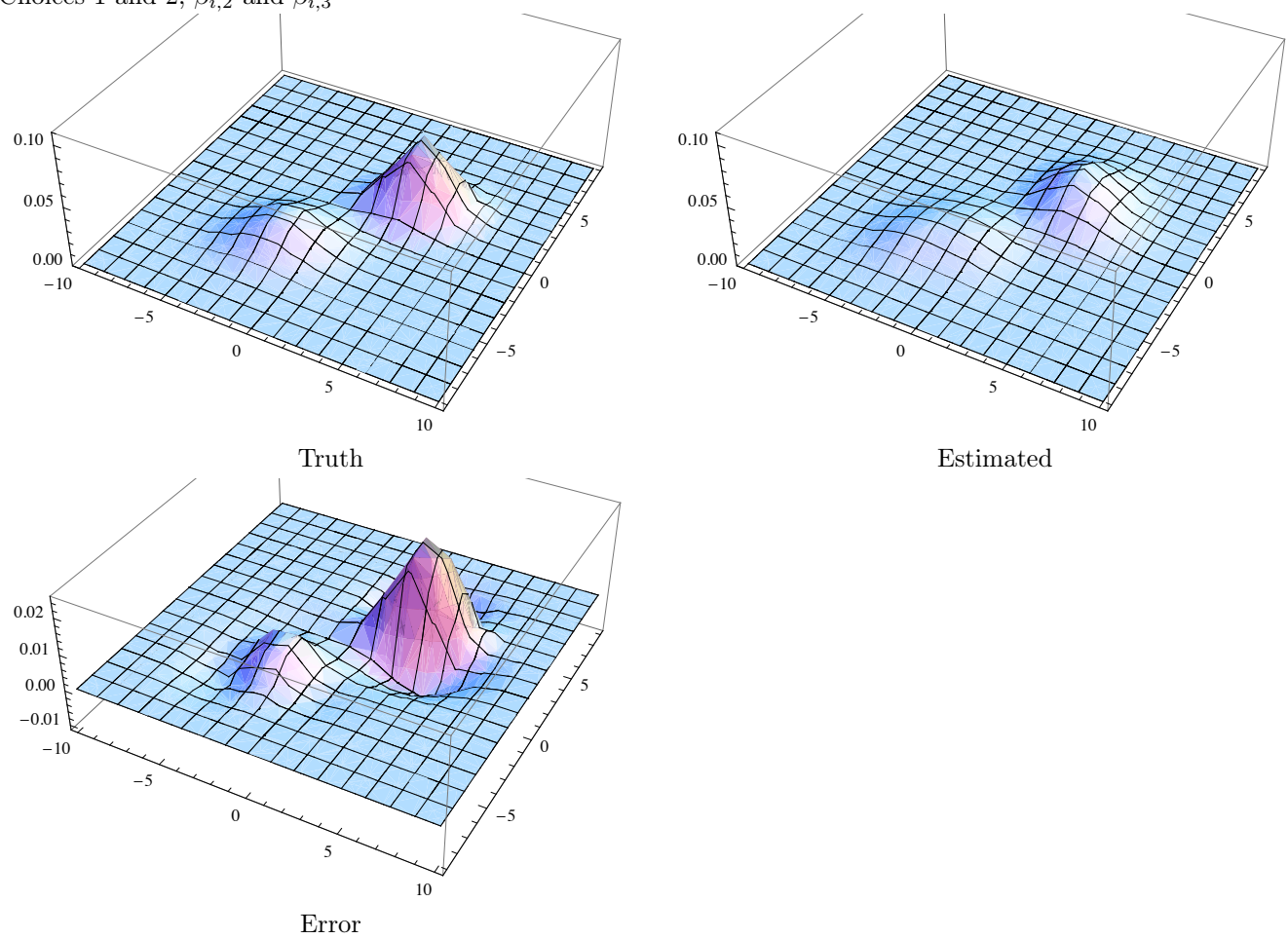
There exist few nonparametric identification theorems for the distribution of heterogeneity in many economic models estimated every day in industrial organization and labor economics. This paper argues that a convenient mathematical formulation of many identification problems is proving linear independence of the space of potential heterogeneous types. However, proving linear independence for each economic model separately can be cumbersome. We introduce an intermediate property of economic models, known as reducibility, that is a sufficient condition for linear independence and hence identification of the distribution of heterogeneity.

We apply reducibility to three classes of models: discrete choice models, continuous choice models, and selection and mixed continuous and discrete choice models. We hope our results place on firmer theoretical ground the wide application of parametric estimators for the heterogeneity distributions in these models. Also, our results open up the use of nonparametric estimators for heterogeneity distributions: identification does not come from assuming a parametric functional form for the heterogeneity distribution. Reducibility can likely be used to show identification of the distribution of heterogeneity for many other economic choice models.

Our results are mathematically general. For many models, we allow the data to enter a nonparametric function $g_\theta(x)$, that is known only to be real analytic. We are nonparametric on the distribution $G(\theta)$ of the functions $g_\theta(x)$. We therefore show identification while working in two infinite-dimensional spaces.

For continuous outcomes, we show identification of the full, joint distribution of heterogeneity in a system of nonparametric, seemingly unrelated regressions. We can also allow endogenous regressors that are determined by an auxiliary equation, as part of a triangular system. We

Figure 5: True and Estimated Joint Density of the Intercepts in the Outcome Equations For Both Choices 1 and 2, $\beta_{i,2}$ and $\beta_{i,3}$



identify the full joint distribution of the nonparametric functions in the equations in the triangular system.

In terms of multinomial choice, relative to the literature we have a least six contributions: 1) we study multinomial choice and not just binary choice, 2) we rely on monotonicity and not the linearity and large support in all characteristics needed to apply the Cramer and Wold theorem, 3) we do not rely on identification at infinity, 4) we identify the joint distribution of product-specific slopes and intercepts for all choices, 5) we are nonparametric on the subutility function $u_{\theta}^j(x_j)$ for choice j , and 6) we give an approach to allowing for endogenous characteristics such as prices.

In terms of selection models, we have four contributions relative to the literature 1) we allow random coefficients in all parts of the model, 2) we identify the joint distribution of the outcomes, 3) we do not rely on identification at infinity, instead using linear independence, and 4) our argument generalizes easily to the case of multinomial choice in the selection equation.

Our linear independence identification strategy, while strictly speaking not constructive, does naturally suggest the linear regression estimator of Bajari, Fox, Kim and Ryan (2007). We perform a fake data exercise using a selection model and show the estimator is capable of recovering the joint distribution of three random coefficients: one random coefficient in the selection equation and two random intercepts in the outcome equations. Subject to regularity conditions, other nonparametric mixtures estimators can be used as well.

References

- Abbring, Jaap H. and Gerard J. van den Berg**, “The Identifiability of the Mixed Proportional Hazards Competing Risks Model,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2003, *65* (3), 701–710.
- Aliprantis, Charalambos D. and Kim C. Border**, *Infinite Dimensional Analysis: A Hitchhiker’s Guide*, third ed., Springer, 2006.
- Andrews, D.W.K. and M.M.A. Schafgans**, “Semiparametric Estimation of the Intercept of a Sample Selection Model,” *Review of Economic Studies*, 1998, *65* (3), 497–517.
- Bach, A., D. Plachky, and W. Thomsen**, “A Characterization of Identifiability of Mixtures of Distributions,” in M. L. Puri, P. Révész, and W. Wertz, eds., *Mathematical Statistics and Probability Theory*, D. Reidel, 1986.
- Bajari, Patrick, Jeremy T. Fox, Kyoo-il Kim, and Stephen Ryan**, “Identification and Estimation of Random Utility Models,” April 2007. working paper.
- Barbe, Philippe**, “Statistical analysis of mixtures and the empirical probability measure,” *Acta Appl. Math.*, 1998, *50* (3), 253–340.

- Beran, R. 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 Models with Heterogeneous Consumers and Endogeneity,” 2007. working paper.
- Blum, JR and V. Susarla**, “Estimation of a Mixing Distribution Function,” *The Annals of Probability*, 1977, 5 (2), 200–209.
- Briesch, Richard A., Pradeep K. Chintagunta, and Rosa L. Matzkin**, “Nonparametric Discrete Choice Models with Unobserved Heterogeneity1,” 2007. working paper.
- Chamberlain, G.**, “Asymptotic Efficiency in Semi-Parametric Models with Censoring,” *Journal of Econometrics*, 1986, 32 (2), 189–218.
- Chernozhukov, V. and C. Hansen**, “An IV Model of Quantile Treatment Effects,” *Econometrica*, 2005, 73 (1), 245–261.
- Cosslett, Stephen R.**, “Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model,” *Econometrica*, 1983, 51 (3), 765–782.
- Cramer, H. and H. Wold**, “Some Theorems on Distribution Functions,” *Journal of the London Mathematical Society*, 1936, 1 (4), 290.
- Day, N.E.**, “Estimating the components of a mixture of normal distributions,” *Biometrika*, 1969, 56 (3), 463.
- Dempster, A.P., N.M. Laird, and D.B. Rubin**, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, 1977, 39 (1), 1–38.
- Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlačil**, “Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects,” *Econometrica*, 2008.
- Gautier, Eric and Yuichi Kitamura**, “Nonparametric Estimation in Random Coefficients Binary Choice Models,” 2007. working paper.
- Gronau, Reuben**, “Wage Comparisons—A Selectivity Bias,” *The Journal of Political Economy*, 1974, 82 (6), 1119–1143.
- Hausman, J. and D. Wise**, “A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences,” *Econometrica*, 1978, 46 (2), 403–426.

- Heckman, J.J.**, “Shadow Prices, Market Wages, and Labor Supply,” *Econometrica*, 1974, *42* (4), 679–694.
- , “Sample Selection Bias as a Specification Error,” *Econometrica*, 1979, *47* (1), 153–161.
- , “Varieties of Selection Bias,” *The American Economic Review*, 1990, *80* (2), 313–318.
- and **B. S. Singer**, “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data,” *Econometrica*, 1984, *52* (2), 271–320.
- and **B.E. Honore**, “The Empirical Content of the Roy Model,” *Econometrica*, 1990, *58* (5), 1121–1149.
- and **E.J. Vytlačil**, “Local instrumental variables and latent variable models for identifying and bounding treatment effects,” *Proceedings of the National Academy of Sciences*, 1999.
- and —, “Instrumental variables, selection models, and tight bounds on the average treatment effect,” *Econometric Evaluation of Labour Market Policies*, 2001.
- and **S. Navarro**, “Dynamic Discrete Choice and Dynamic Treatment Effects,” *Journal of Econometrics*, 2007.
- , **J. Smith**, and **N. Clements**, “Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts,” *Review of Economic Studies*, 1997, *64* (4), 487–535.
- , **S. Urzua**, and **E. Vytlačil**, “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, 2006, *88* (3), 389–432.
- Hoderlein, Stefan, Jussi Klemelä, and Enno Mammen**, “Analyzing the Random Coefficient Model Nonparametrically,” June 2008. working paper.
- Hong, Han and Elie Tamer**, “Endogenous binary choice model with median restrictions,” *Economics Letters*, 2003, *80*, 219–225.
- 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, G.W. and J.D. Angrist**, “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 1994, *62* (2), 467–475.
- Khan, Shakeeb and Elie Tamer**, “Irregular Identification, Support Conditions and Inverse Weight Estimation,” 2007. working paper.

- Laird, Nan**, “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution,” *Journal of the American Statistical Association*, 1978, 73 (364), 805–811.
- Lee, Lung-Fei**, “Semiparametric maximum likelihood estimation of polychotomous and sequential choice models,” *Journal of Econometrics*, 1995, 65, 381–428.
- Lewbel, Arthur**, “Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables,” *Journal of Econometrics*, 2000, 97 (1), 145–177.
- Li, J.Q. and A.R. Barron**, “Mixture density estimation,” *Advances in Neural Information Processing Systems*, 2000, 12, 279–285.
- Lindsay, Bruce G. and Katherine Roeder**, “Uniqueness of Estimation and Identifiability in Mixture Models,” *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 1993, 21 (2), 139–147.
- Manski, C.F.**, “Maximum Score Estimation of the Stochastic Utility Model of Choice,” *Journal of Econometrics*, 1975, 3 (3), 205–228.
- and **J.V. Pepper**, “Monotone Instrumental Variables: With an Application to the Returns to Schooling,” *Econometrica*, 2000, 68 (4), 997–1010.
- Matzkin, Rosa L.**, “Nonparametric identification and estimation of polychotomous choice models,” *Journal of Econometrics*, 1993, 58, 137–168.
- , “Heterogeneous Choice,” 2007. working paper.
- McFadden, Daniel L. and Kenneth Train**, “Mixed MNL Models for Discrete Response,” *Journal of Applied Econometrics*, 2000, 15, 447–470.
- Narasimhan, Raghavan**, *Several Complex Variables*, University of Chicago Press, 1971.
- Rossi, P.E. and G.M. Allenby**, “Bayesian Statistics and Marketing,” *Marketing Science*, 2003, 22 (3), 304–328.
- , **G.M. Allenby**, and **R. McCulloch**, *Bayesian Statistics and Marketing*, John Wiley and Sons, 2005.
- Roueff, Francois and Tobias Rydén**, “Nonparametric estimation of mixing densities for discrete distributions,” *The Annals of Statistics*, 2005, 33 (5), 2066–2108.
- Shaikh, Azeem M. and Edward Vytlačil**, “Threshold Crossing Models and Bounds on Treatment Effects: A Nonparametric Analysis,” 2005. working paper.
- Sørensen, Morten**, “Identification of General Selection Models,” May 2006. working paper.

- Teicher, H.**, “Identifiability of Mixtures,” *The Annals of Mathematical Statistics*, 1961, 32 (1), 244–248.
- 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.
- Train, Kenneth**, “EM Algorithms for Nonparametric Estimation of Mixing Distributions,” *Journal of Choice Modeling*, 2008.
- Vytlacil, Edward and Nese Yildiz**, “Dummy Endogenous Variables in Weakly Separable Models,” *Econometrica*, 2007.
- Wooldridge, Jeffrey M.**, “1 2 Instrumental Variables Estimation of the Average Treatment Effect in Cor- Instrumental Variables Estimation of the Average Treatment Effect in Correlated Random Coefficient Models,” in D. Milliment, J. Smith, and E. Vytlacil, eds., *Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics*, Vol. 21, Elsevier, 2007.
- Yakowitz, Sidney J. and J. Sprangins**, “On the Identifiability of Finite Mixtures,” *The Annals of Mathematical Statistics*, 1968, 39, 209–214.