

Identifying Heterogeneity in Economic Choice and Selection Models Using Mixtures

Jeremy T. Fox

University of Chicago and NBER

Amit Gandhi

University of Wisconsin*

March 2009

Abstract

We show how to nonparametrically identify the distribution of heterogeneity in a general class of structural economic choice models. We state an economic property known as reducibility and prove that reducibility ensures identification. Reducibility makes verifying the identification of nonlinear models a straightforward task because it is a condition that is stated directly in terms of a choice model. We can allow for a nonparametric distribution over nonparametric functions of the data. We use our framework to prove identification in three classes of economic models: 1) nonparametric regressions including with endogenous regressors, 2) multinomial discrete choice including endogenous regressors as well as multiple purchases with complementarities, and 3) selection and mixed continuous-discrete choice. Our identification strategy avoids 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.

*Thanks to Steven Durlauf, James Heckman, Salvador Navarro, Philip Reny, Azeem Shaikh, Morten Sørensen and Edward Vytlačil for helpful comments. Also thanks to seminar participants at Brown, Caltech, CREST, Chicago, Cowles, EC2 Rome, LSE, Michigan State, Rochester, Stanford, Toulouse, USC 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.

1 Introduction

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. More generally, 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 approach to 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. Finally, showing that a complex economic model is identified makes empirical researchers more comfortable with estimates from the model.

The key advantage of our approach to establishing identification is that we develop an identification condition that is expressed directly in terms of the underlying economic choice model of interest. We use the term “choice model” in a broad sense as a term for any model that specifies the response of an agent with certain characteristics to an economic environment with specified characteristics. If we denote the characteristics of the agent by θ and the economic environment by x , then the model is given by $y = f_\theta(x)$, where y is the response. While the econometrician can observe x , the agent’s characteristics θ are unobservable and heterogeneous in the underlying population of agents. A key feature of the model is that knowledge of the distribution of the unobservable characteristics θ is essential for answering particular economic questions, and it is this distribution that constitutes the target of inference.

Different choice models are generated by different assumptions about the form of the response y . We focus attention 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 endogenous regressors, 3) linear simultaneous equations models with all parameters being random across markets, 4) the multinomial choice demand model, including extensions that allow for endogenous regressors and multiple purchases with complementarities across products, 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). Before presenting a more detailed outline of our main results, we first establish the general nature of the identification problem as it applies to our setting.

2 The Identification Problem

We consider a general class of economic models that can be represented as a relationship $y = f_\theta(x)$, where $y \in \mathbb{R}^m$ is the response of an agent of type $\theta \in \Theta$ to the economic environment $x \in \mathcal{X}$. The model is thus described by a triple $\mathcal{M} = (f, \mathcal{X}, \Theta)$, 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 product characteristics $x_j \in \mathbb{R}^K$ and an additional scalar characteristic w_j .¹ Each agent has a type $\theta \in \Theta$ that characterizes its preferences, and thus an agent of type θ has a utility function $u_\theta^j(x) + w_j$ for product j . Faced with the choice set x , a type θ agent chooses the alternative $j \in \{1, \dots, J\}$ if $u_\theta^j(x) + w_j \geq u_\theta^k(x) + w_k$ for all $k \neq j$.²

We pursue an approach to identification that does not place any structure on the type space Θ , and in particular we allow it to be an infinite dimensional space. In the random utility model, θ could index a J -tuple of choice-specific utility functions $\{u_\theta^j(x)\}_{j=1}^J$ and Θ could denote the set of all such J -tuples of utility functions where each utility function $u_\theta^j(x)$ is restricted to only satisfy nonparametric regularity conditions.³ More generally, we will speak of Θ as the set of possible types of agents, and we will not require any a priori structure on the this space.

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 over Θ 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. This is

¹Typically price can play the role of the scalar.

²We will not need an additive term such as w for nonparametric regression. Also, we will investigate the pure characteristics demand model to show that our identification strategy for multinomial choice does not rely critically on large support of w .

³The main regularity condition we impose on real valued functions over real vectors x is that they are real analytic, which we define rigorously in section 4.

because each $u_\theta^j(x)$ is an actual function that gives the subutility of type θ for choice j at all $x \in \mathcal{X}$. 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, i.e., they do not vary within the population.

The data consist of observations $\{(y_i, x_i)\}_{i=1}^N$ on the underlying population. The econometrician has no special knowledge about each unit of observation's θ_i , other than the outcome y_i and environment x_i . From the observables, the econometrician can identify the population parameter $F(y | x)$, the conditional distribution function of the response, for any $y \in \mathbb{R}^m$ 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.⁴ The question is whether the distribution over types G can be identified from the variation X available in the data. As the variation X plays a critical role in identification, we will denote the full model by the pair (\mathcal{M}, X) .

The assumption that the distribution over types G is stable across economic environments $x \in X$ implies that θ is stochastically independent of x , allowing us to express

$$F(y | x) = G(\{\theta \in \Theta \mid f_\theta(x) \leq y\}) = \int 1[f_\theta(x) \leq y] dG(\theta). \quad (1)$$

Thus we immediately see that G is identified up to the measure it assigns to sets of the form $I_{y,x}^\Theta = \{\theta \in \Theta \mid f_\theta(x) \leq y\}$. The problem is whether the class of sets

$$\mathcal{I}^\Theta = \{I_{y,x}^\Theta \mid y \in \mathbb{R}^m, x \in X\} \quad (2)$$

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 functions $F(y | x)$ defined over $\mathbb{R}^m \times X$. 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 (\mathcal{M}, X) is *identified* relative to \mathcal{G} if L is one-to-one. That is, if $G, G' \in \mathcal{G}$, $G \neq G'$, then there exists an experiment in the data $(y, x) \in \mathbb{R}^m \times X$ such that $F_G(y | x) \neq F_{G'}(y | x)$.⁵

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}, X) that ensures the exis-

⁴Mathematically, our ability to predict out of sample (from X to all of \mathcal{X}) comes from the restriction to real analytic functions, as we will show. More generally, out of sample predictions could also be possible if the researcher imposed functional forms which are motivated by theory and are special cases of our real analytic class.

⁵In Appendix A, we discuss extending this definition of identification to require that a positive probability of such distinguishing experiments x exists. Identification with a positive probability is easy to verify for the models we study in this paper once their identification under the definition presented in the main text has been established.

tence of such an experiment, and hence identification. We term this condition reducibility. The main idea behind reducibility can be explained by considering three types θ_1 , θ_2 , and θ_3 making a binary choice between outcomes 1 and 2. Say at some budget set x types θ_1 and θ_2 pick 2 and type θ_3 picks 1. The model is reducible if we can find some new budget set x' where one of θ_1 and θ_2 , say θ_1 , still picks 2 and the other switches to choice 1 and θ_3 still picks 1. Thus, we have reduced the set of types picking 2 from $\{\theta_1, \theta_2\}$ to $\{\theta_1\}$.

We show that reducibility allows us to identify the distribution G over types in a fully non-parametric fashion. We take full nonparametric identification to mean not putting any parametric structures on either the type space Θ or the distribution G . Thus agents can be indexed by types θ that index elements of an infinite-dimensional space Θ , and any potential distribution G over types lies an infinite dimensional family of distributions \mathcal{G} . A lack of nonparametric identification calls into question any parametric estimator of the model: apparently the parametric estimator is only consistent because of parametric functional form restrictions on the response model $f_\theta(x)$ or parametric or finite-dimensional restrictions on the class \mathcal{G} .

To achieve such generality, we must impose some nonparametric regularity conditions on the problem. In particular, we restrict attention to a particular nonparametric class of distributions \mathcal{G} that we argue is essentially without loss of generality for empirical practitioners as a way to model heterogeneity in economic models. This is the class \mathcal{G} of all multinomial distributions over Θ . Thus the only restriction being placed on the distribution of types $G \in \mathcal{G}$ is that the set of types having positive support in the population is finite. However the number of support points, the location of the support points, and their masses are a priori unknown and need to be identified from the data. Thus \mathcal{G} constitutes an infinite dimensional space of distributions.⁶ Furthermore, the class \mathcal{G} is defined without requiring any structure on Θ . Hence we are able to be fully nonparametric about the type space Θ , thus allowing us to build a general theory.⁷

We argue that the class of distributions \mathcal{G} is an especially natural assumption in the context of economic choice models. The most basic defense is that even if every individual had its own type, the type space would be ultimately finite as real economic populations are inherently finite, although large and potentially complicated. While arbitrarily complicated discrete distributions are an inconvenient assumption for empirical work, and hence the widespread adoption of continuous distributions that are capable of approximating them in estimation contexts, a solution to the identification problem should not rely upon the abstraction that continuous measures impose

⁶The class of multinomial distributions \mathcal{G} over any infinite set Θ is an infinite-dimensional space. Assume to the contrary that the space \mathcal{G} was instead k -dimensional for a finite integer k . Then any $k + 1$ elements of \mathcal{G} would be linearly dependent. Let δ_θ denote the Dirac delta probability measure that assigns mass 1 to $\theta \in \Theta$. Because Θ is an infinite set, we can always find $k + 1$ elements of Θ , say $\{\theta_1, \dots, \theta_{k+1}\}$, and as a result we can always find $k + 1$ elements of \mathcal{G} , namely $\{\delta_{\theta_1}, \dots, \delta_{\theta_{k+1}}\}$. However $\{\delta_{\theta_1}, \dots, \delta_{\theta_{k+1}}\}$ can never be a linearly dependent set. Thus \mathcal{G} must be infinite dimensional.

⁷This contrasts with the non-nested 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-dimensional spaces.

upon finite populations. The critical question behind the identification problem is whether the same economic population G facing exogenously varying economic environments $x \in X$ will have revealed preferences in the form of the reduced form relationship in the data $F(y | x)$ that are informative enough to identify G . To assume that the population G is finite is an inherently un-falsifiable assumption as 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. Thus assuming that $G \in \mathcal{G}$ is simply a nonparametric regularity condition (much like other technical regularity conditions such as continuity or smoothness in other identification contexts) that does not restrict in any way a priori the potential realizations of any data set. 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 (Aliprantis and Border, 2006, Theorem 15.10). Hence any policy question can be addressed to an arbitrary degree of precision by the class \mathcal{G} . Thus for empirical work, \mathcal{G} is practically 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).⁸ Bajari, Fox, Kim and Ryan (2009) 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 3 presents our identification results for generic economic choice models using reducibility. We then show that reducibility is satisfied in a class of important structural models used in applied microeconomics. Section 4 considers continuous outcome models, including the nonparametric regression model with endogenous regressors. Section 5 considers multinomial choice models, including those with price endogeneity and complementarities across multiple products. 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

⁸Another 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).

specific class of model. Section 7 motivates an estimator for $G \in \mathcal{G}$ that arises from our identification logic, and presents a fake data experiment using this estimator for selection models.

3 Identification Using Reducibility

Recall the basic question is whether the class of sets \mathcal{I}^Θ (as defined by (2)) generated by the model (\mathcal{M}, X) is rich enough to identify G within the class of distributions \mathcal{G} . We now show that an affirmative answer to this question holds under a natural condition on (\mathcal{M}, X) that we term reducibility. We first define a key concept.

Definition 3.1 (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*

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

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 T^0 that corresponds to the support of the true underlying population distribution $G^0 \in \mathcal{G}$. The key feature of I-sets is that they are strictly a property of the underlying economic choice 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 full model (\mathcal{M}, X) is capable of generating sufficient variation in I-sets. To formalize this condition, let \mathcal{I}^T be the class of all I-sets for any finite subset of types $T \subset \Theta$, $\mathcal{I}^T = \{I_{y,x}^T \mid y \in \mathbb{R}^m, x \in X\}$. Our axiom for nonparametric identification is the following.

Definition 3.2 (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 contained in $I_{y,x}^T$.*
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 contained in $I_{y,x}^T$ as reducing the I-set $I_{y,x}^T$. Reducibility has the following simple but useful implication.

Lemma 3.1. *If the model (\mathcal{M}, X) is reducible, then for any finite $T \subset \Theta$, there exists a singleton $I_{y,x}^T \in \mathcal{I}^T$.*

Proof. Take a non-empty $I_{y,x}^T \in \mathbb{I}^T$. If it is a singleton, then we are done. If not, then reduce $I_{y,x}^T$ to $I_{y',x'}^T \in \mathbb{I}^T$, and repeat the argument. After a finite number of steps we will have produced a singleton I-set. □

We now state and prove our main result.

Theorem 3.1. *If the model (\mathcal{M}, X) is reducible, then the model is identifiable with respect to \mathcal{G} .*

Proof. Recall that identification requires showing that the mapping $L : \mathcal{G} \rightarrow \mathcal{F}$ defined by (1) is one to one. Thus for $G^0, G^1 \in \mathcal{G}$ with $G^0 \neq G^1$, we must have that $F_{G^0}(y | x) \neq F_{G^1}(y | x)$ for some $(y, x) \in \mathbb{R}^m \times X$. To show that L is one to one, we take a point $F \in L(\mathcal{G})$ and show $L(G^0) = L(G^1) = F$ implies $G^0 = G^1$.

Observe that we can represent any $G \in \mathcal{G}$ by a pair (T, p) , where the probability vector $p = \{p_\theta\}_{\theta \in T} \in \Delta^{n-1}$ puts non-negative mass over a finite set of types $T = \{\theta_i\}_{i=1}^n \subset \Theta$ for some positive integer n . Thus for $G \in \mathcal{G}$, we can express (1) as

$$F(y | x) = \sum_{\theta \in I_{y,x}^T} p_\theta.$$

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

Assume to the contrary that $\pi_\theta \neq 0$ for some $\theta \in T$. Since we have that for all $(y, x) \in \mathbb{R}^m \times X$, $F_{G^0}(y | x) = F_{G^1}(y | x) = F(y | x)$, it follows that

$$\sum_{\theta \in I_{y,x}^T} \pi_\theta = 0, \tag{3}$$

for all I -sets $I_{y,x}^T \in \mathcal{I}^T$.

Now let $T^2 = \{\theta \in T \mid \pi_\theta \neq 0\}$, which by the assumption that $G^0 \neq G^1$ is non-empty. Then applying reducibility and so Lemma 3.1, we can produce a singleton $I_{y,x}^{T^2} = \{\theta^*\}$. Furthermore, we can re-write (3) as

$$\sum_{\theta \in I_{y,x}^T} \pi_\theta = \sum_{\theta \in I_{y,x}^{T^2}} \pi_\theta + \sum_{\theta \in I_{y,x}^{T-T^2}} \pi_\theta = \sum_{\theta \in I_{y,x}^{T^2}} \pi_\theta = \pi_{\theta^*} = 0 = \pi_{\theta^*} \neq 0.$$

This is a contradiction because it contradicts the construction that $\theta^* \in T^2$. Hence it must be that $\pi_\theta = 0$ for all $\theta \in T$, which implies $G^0 = G^1$ and hence identification. \square

It is critical to understand that the act of reducing an I -set has nothing to do with finding the actual experiment $(y, x) \in \mathbb{R}^m \times X$ that separately identifies G^0 from G^1 , i.e., the experiment (y, x) for which $F_{G^0}(y | x) \neq F_{G^1}(y | x)$. That is, our main theorem is properly viewed as an existence theorem, and asserts that under reducibility of the model, such an experiment must always exist. It

is non-constructive in the sense that it does not inform us about how to find this particular (y, x) . The singleton sets that were generated in the proof are independent of the process of finding an actual experiment in the data that distinguishes G^0 and G^1 .

Said another way, all variation in $x \in X$ is potentially “identifying” because such variation translates into variation in I -sets $I_{y,x}^T$, which is the key to showing that a model is reducible. 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.

Consider applying Theorem 3.1 to two simple choice models, purely to illustrate the verification of reducibility. First, consider a model without covariates in which the outcome $f(\theta)$ is a known function that is strictly increasing in an agent’s type $\theta \in \mathbb{R}$. Let $T = \{\theta_1, \theta_2, \theta_3\}$ and $I_y^T = \{\theta \in T \mid f(\theta) \leq y\} = \{\theta_1, \theta_2\}$. As $f(\theta_1) \neq f(\theta_2)$ because f is strictly increasing, we can set $y' = \min\{f(\theta_1), f(\theta_2)\} = f(\theta_1)$, say. Then $I_{y'}^T = \{\theta_1\} \subset \{\theta_1, \theta_2\} = I_y^T$ and $I_{y'}^T \neq \emptyset$, so the model is reducible and hence the distribution of types G is identified. For an example where reducibility fails, consider the model where each type $\theta = (\theta^a, \theta^b)$ is a vector of two real-valued scalars, and where the outcome is $f(\theta) = \theta^a + \theta^b$. Again let $T = \{\theta_1, \theta_2, \theta_3\}$ and assume these types are such that $f(\theta_1) = \theta_1^a + \theta_1^b = f(\theta_2) = \theta_2^a + \theta_2^b$. Then it is not possible for an I -set I_y^T to contain θ_1 and not θ_2 , or vice versa. Thus the model is not reducible, which is consistent with the fact that this model is clearly not identified.

This paper will be concerned with the identification of heterogeneity in a series of related economic choice models that are widely used in applied microeconomics. All of our proofs for identification are based on showing reducibility of the model. As different economic choice models generate different structures on the nature of the outcome variable, it is convenient to define an I -set in a more model-specific way. There are two main special cases of the general model, which are shown below. For each case, we provide an alternative definition of an I -set that is compatible with the proof of Theorem 3.1.

- Discrete response: $j = f_\theta(x)$ for a discrete valued $j \in J$. 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

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

- Mixed Discrete / Continuous response: $(y, j) = (f_\theta^1(x), f_\theta^2(x))$ for a continuous $y \in \mathbb{R}^m$ and a discrete $j \in J$. 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

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

While we have defended the class of distributions \mathcal{G} on the grounds of its sufficient generality, the ideas behind reducibility can also be applied if we impose the alternative restriction that every $G \in \mathcal{G}$ admits a density function. This is discussed in Appendix B. It is important to observe that the class of distributions that admit a density function is not more general than the class of multinomial distributions. We provide the argument in Appendix B only to show robustness of the reducibility concept.

While reducibility is sufficient for identification, we have not claimed that it is necessary. Teicher (1963) and Yakowitz and Spragins (1968) investigate the identification of finite mixtures in statistical models without covariates. They show that a necessary and sufficient condition for identification is that the statistical model satisfies a property known as linear independence. However in the context of an economic choice model, linear independence is a non-primitive assumption on the model, and thus showing linear independence of (\mathcal{M}, X) would be equivalent to showing identification itself, leaving us back where we started. The key contribution of reducibility is that it is expressed in terms of an economic choice model and thus can be verified on the basis of the underlying behavior of the agents in the model (\mathcal{M}, X) , a point that will become more clear in subsequent sections.⁹

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 $y = f_\theta(x)$ 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 $f_\theta(x) = f(x_1, \theta_1, \alpha(x_2, \theta_2))$, 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

⁹Blum and Susarla (1977) and Bach, Plachky and Thomsen (1986) have extended work on linear independence and finite mixtures to, respectively, the non-nested class of distributions that admit a density and the class of all distributions.

decomposition, then we can prove identification of the underlying distribution over types $G(\theta)$ by a two step procedure, which we now establish.

Theorem 3.2. *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.*

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

4 Nonparametric Regression

We first focus attention on identifying heterogeneity in the nonparametric regression model. A type θ indexes a causal relationship $y = f_\theta(x)$ for $x \in \mathcal{X} = \mathbb{R}^K$ and $y \in \mathbb{R}^m$. Hence a type θ indexes a function $f_\theta : \mathbb{R}^K \rightarrow \mathbb{R}^m$, and the type space Θ denotes the set of all such functions subject to regularity conditions. We allow vectors of functions and outcomes. We do not impose any parametric structure on a type θ , and hence Θ will constitute an infinite dimensional space of functions.

Observe that in the conventional language of econometrics, x plays the role of the regressors and θ plays the role of the econometric error in the model. Thus the assumption that the population distribution G over Θ is stable across economic environments x is equivalent to assuming independence between the econometric error θ and the regressors x . The linear regression model with random coefficients is the simplest special case of the model.

There are two main regularity conditions that we will require. The first is that the support $X \subseteq \mathbb{R}^K$ of the regressors contains an open subset. As we will only use variation in x within such an open subset, it is without loss to assume simply that X is itself open. Observe that this support requirement does not place any restriction on the “size” of X . In particular, X can be contained in an arbitrarily small ball.

However the assumption that X is open does rule out the possibility of regressors that only admit discrete variation. Nevertheless, if there existed any such discrete regressors r , we can always condition on them throughout the analysis. For example, we can set up the problem as one of fixing the value of r and identifying the distribution $G(\theta | r)$ over functions $f_\theta(x; r)$ using just open set variation in x , and repeat the identification argument for every discrete value of r . We shall implicitly condition on discrete regressors in this fashion for the remainder of the paper and thus ignore any further considerations about discrete regressors.

The second regularity condition we will require is that for any type $f_\theta(x) = (f_\theta^1(x), \dots, f_\theta^m(x))$, each component function $f_\theta^i(x)$ for $i = 1, \dots, m$ is a real analytic function.¹⁰ Hence the type space

¹⁰Let X be a non-empty, open subset of \mathbb{R}^K . Following Abbring and van den Berg (2003), a function $f : \mathbb{R}^K \rightarrow \mathbb{R}^m$

Θ can be defined as the set of all mappings $f_\theta : \mathbb{R}^K \rightarrow \mathbb{R}^m$ with component functions that are real analytic. A fundamental property of real analytic functions that we shall exploit is the fact that for any two distinct real analytic functions $f, f' : \mathbb{R}^K \rightarrow \mathbb{R}$, and for any open, connected set $U \subseteq \mathbb{R}^K$, f and f' cannot agree on the whole of U : there must exist $x \in U$ for which $f(x) \neq f'(x)$ (Krantz and Parks, 2002, Corollary 1.2.6). This property plays an essential role in showing reducibility. If a function $f_\theta : \mathbb{R}^K \rightarrow \mathbb{R}^m$ is such that each of its m component functions is real analytic, then we shall say that the f_θ is vector analytic. It is straightforward to see that vector analytic functions inherit the preceding property of real analytic functions, namely that for any two distinct vector analytic functions $f, f' : \mathbb{R}^K \rightarrow \mathbb{R}^m$, and for any open, connected set $U \subseteq \mathbb{R}^K$, f and f' cannot agree on the whole of U .

Theorem 4.1. *The nonparametric regression model is identified with respect to \mathcal{G} .*

Proof. We show identification by showing reducibility of the model. For any finite set of types $T \subset \Theta$, consider any non-empty I -set $I_{y,x}^T = \{\theta \in T \mid f_\theta(x) \leq y\}$. 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. Without loss of generality, we can restrict attention to the case where $T = I_{y,x}^T \cup T_3$ and $I_{y,x}^T = \{\theta_1, \theta_2\}$. We only need to show we can reduce $I_{y,x}^T$ by eliminating one agent, so disregarding other agents in $I_{y,x}^T$ is without loss of generality.

First consider the case where $y_{\theta_1} = f_{\theta_1}(x) \neq y_{\theta_2} = f_{\theta_2}(x)$. Each y may be a vector, and so there must be a scalar component k where, say, $y_{\theta_1}^k < y_{\theta_2}^k$. Set $y' = y_{\theta_1}$. By the properties of a partial order, $y_{\theta_2} \not\leq y'$. For $\theta_3 \in T_3$, $\theta_3 \notin I_{y,x}^T$, $f_{\theta_3}(x) \not\leq y$ and so $f_{\theta_3}(x) \not\leq y'$. Therefore, $I_{y',x}^T = \{\theta_1\} \subset I_{y,x}^T$ and $I_{y',x}^T \neq \emptyset$.

Now consider the case where $y_{\theta_1} = f_{\theta_1}(x) = y_{\theta_2} = f_{\theta_2}(x)$. Observe that because T is finite, we can raise y by a sufficiently small amount to y' so that $I_{y,x}^T = I_{y',x}^T$ (no types exit or enter the I -set). This gives us that for each $\theta \in I_{y',x}^T$, $f_\theta(x) < y'$, and for each $\theta \in T_3$, $f_\theta^k(x) > y'_k$ for some $k \in \{1, \dots, M\}$. By the continuity of each type's f_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, by the property of vector analytic functions discussed above we can pick an $x' \in U$ so that $f_{\theta_1}(x') \neq f_{\theta_2}(x')$. Now suppose $f_{\theta_1}^k(x') < f_{\theta_2}^k(x')$ for some component $k \in \{1, \dots, M\}$. By a similar argument to the previous paragraph, we can set $y'' = f_{\theta_1}(x')$ and show that $I_{y'',x'}^T$ is a non-empty set that reduces $I_{y,x}^T$. \square

4.1 Endogenous Regressors Through a Triangular System

Endogenous regressors are often encountered in social-science applications. Thus in the context of the model $f_\theta : \mathbb{R}^K \rightarrow \mathbb{R}^M$, it is possible that some subset of the regressors, say the first $J < K$ is real analytic if, given $\xi \in X$, there is a power series in $x - \xi$ that converges to $f(x)$ for all x in some neighborhood $U \subset \mathbb{R}^k$ of ξ . Real analytic functions must be smooth.

regressors, are 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 around” the endogenous regressors according to $\tilde{x} = g_\theta(x, z)$.¹¹ Thus the model can be expressed as a recursive system of equations

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

and a type θ thus indexes a pair of vector-valued functions (f_θ, g_θ) . 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 distribution of the causal relationship $y = f_\theta(\tilde{x}, x)$. In particular, if the distribution of types (f_θ, g_θ) can be recovered, then we can recover the distribution of the causal effect $\frac{\partial}{\partial \tilde{x}} f_\theta(\tilde{x}, x)$, which in many cases is the main structural feature of interest.

A special case of this model is linear 2SLS with all of the coefficients being random. Indeed, the random coefficients in the first stage have an unrestricted joint distribution with the random coefficients in the second stage. Let z be a tuition subsidy, \tilde{x} the amount of schooling completed and y wages. Our model allows both sorting on the gain from schooling (schooling \tilde{x} is correlated with the random coefficient on \tilde{x} in the outcome equation) and for a new phenomenon in the empirical literature: the response to the tuition subsidy may be greater for those whose wages are most sensitive to schooling. This heterogeneity in the response to the instrument and its joint dependence with heterogeneity in response to the endogenous regressor has often been overlooked. Further, our model lacks any monotonicity restrictions. Through a general equilibrium effect, some people might actually reduce schooling in response to a subsidy. For example, wealthy but low-ability students might reduce schooling after the increased enrollment from the subsidy raises the level of competition for jobs that require advanced schooling.

We will show nonparametric identification of heterogeneity so long as the instruments satisfy the condition of being locally valid instruments, which we now define. 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_θ, g_θ) where f_θ and g_θ are vector analytic and where, for each $x \in X$, the derivative matrix $D_z g_\theta(z, x)$ with respect to z has full rank J for almost all (in the sense of Lebesgue measure) $z \in Z$. Such a full rank restriction is a formal way of saying that the instrument z is a locally valid instrument almost everywhere: for any type θ , local variation 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.

¹¹We 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.

Thus fixing $x \in X$ and for almost all $z \in Z$, 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 4.2. *The model with endogenous regressors 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

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

Without loss of generality, we can restrict attention to the case where $T = I_{y, \tilde{x}, x, z}^T \cup T_3$ and $I_{y, \tilde{x}, x, z}^T = \{\theta_1, \theta_2\}$. We proceed by showing reducibility.

First assume $g_{\theta_1} \neq g_{\theta_2}$ as functions. Then apply the reducibility argument from the proof of Theorem 4.1 to show reducibility.

If $g_{\theta_1} = g_{\theta_2} = g_\theta$, then $f_{\theta_1} \neq f_{\theta_2}$ as the two types are assumed to be different. Following a similar logic as the proof of Theorem 4.1, we can raise (y, \tilde{x}) by a sufficiently small amount to (y', \tilde{x}') so that $I_{y', \tilde{x}', x, z}^T = I_{y, \tilde{x}, x, z}^T$. Moreover by continuity of (f_θ, g_θ) for each $\theta \in T$, there exists an open set $V_1 \times V_2 \subset X \times Z$ containing (x, z) such that for each $\theta \in T$, $I_{y', \tilde{x}', x', z'}^T = I_{y', \tilde{x}', x, z}^T$ for all $(x', z') \in V_1 \times V_2$. By the regularity conditions on the instrument, we can find a $z^* \in V_2$ such that $D_z g_\theta(z^*, x)$ has full rank, and given that vector analytic functions are continuously differentiable, there exists an open neighborhood $W_1 \times W_2 \subset V_1 \times V_2$ containing (x, z^*) such that for all $(x, z) \in W_1 \times W_2$, $D_z g_\theta(z, x)$ has full rank J . Hence the change of variable mapping $(x, z) \mapsto (g(x, z), x)$ (which we can denote as L) over $W_1 \times W_2$ is an open mapping by consequence of the open mapping theorem.¹² Now using the fact that both f_{θ_1} and f_{θ_2} are vector analytic, there exists $(x', z') \in W_1 \times W_2$ such that $f_{\theta_1}(g_\theta(x', z'), x') \neq f_{\theta_2}(g_\theta(x', z'), x')$ (this follows more precisely because $L(W_1 \times W_2)$ is an open set in \mathbb{R}^{J+N}). Suppose without loss of generality that $f_{\theta_1}^k(g_\theta(x', z'), x') < f_{\theta_2}^k(g_\theta(x', z'), x')$ for some component $k \in \{1, \dots, M\}$. Letting $y'' = f_{\theta_1}^k(g_\theta(x', z'), x')$ and $\tilde{x}'' = g_\theta(x', z')$, we have that $I_{y'', \tilde{x}'', x', z'}^T$ reduces $I_{y, \tilde{x}, x, z}^T$. \square

4.2 Linear Simultaneous Equations

Simultaneous equations arise when two groups of agents, say firms and consumers, interact to create the endogenous variables. The standard example in economics is that suppliers and demanders

¹²The matrix of partial derivatives of L is of the form $A = \begin{bmatrix} D_x g_\theta(z, x) & I_N \\ D_z g_\theta(z, x) & 0_{J,N} \end{bmatrix}$, where I_N is an identity matrix with N rows and $0_{J,N}$ is a matrix of all 0's with J rows and N columns. The matrix A is invertible because $D_z g_\theta(z, x)$ is invertible. Therefore, by the open-mapping theorem, $(x, z) \mapsto (g(x, z), x)$ is an open mapping.

interact in market equilibrium, where supply equals demand. Let the model be

$$\begin{aligned} q_\theta^d(x) &= \alpha_{0,\theta} + \alpha_{1,\theta}x_1^d + \dots + \alpha_{K,\theta}x_K^d + \gamma_\theta^d p_\theta(x) \\ q_\theta^s(x) &= \beta_{0,\theta} + \beta_{1,\theta}x_1^s + \dots + \beta_{L,\theta}x_L^s + \gamma_\theta^s p_\theta(x) \\ q_\theta^d(x) &= q_\theta^s(x), \end{aligned}$$

where the type θ can be thought of as a market, $q_\theta^d(x)$ is the quantity demanded in market θ at price $p_\theta(x)$, $q_\theta^s(x)$ is the quantity supplied in market θ at price $p_\theta(x)$, and in equilibrium supply equals demand, so $q_\theta^d(x) = q_\theta^s(x)$. The K demand shifters are $x^d = (x_1^d, \dots, x_K^d)$ and the L supply shifters are $x^s = (x_1^s, \dots, x_L^s)$ and $x = (x^d, x^s)$.¹³ A type θ of a market indexes all the parameters

$$(\alpha_{0,\theta}, \alpha_{1,\theta}, \dots, \alpha_{K,\theta}, \gamma_\theta^d, \beta_{0,\theta}, \beta_{1,\theta}, \dots, \beta_{L,\theta}, \gamma_\theta^s).$$

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_θ^d and γ_θ^s . Thus, this model allows unobserved randomness (demand or supply shocks indexed by θ) to not only shift the supply and demand curves across markets, as in the standard textbook model, but to also rotate the supply and demand curves across markets. Different markets have different slopes of the supply and demand curves. Under standard economic assumptions, $\gamma_\theta^d < 0$ and $\gamma_\theta^s > 0$: 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_θ, q_θ) is a function of all the random coefficients. To see this, solve for the reduced forms of price and quantity,

$$\begin{aligned} p_\theta(x) &= \frac{1}{\gamma_\theta^s - \gamma_\theta^d} (\alpha_{0,\theta} + \alpha_{1,\theta}x_1^d + \dots + \alpha_{K,\theta}x_K^d) - \frac{1}{\gamma_\theta^s - \gamma_\theta^d} (\beta_{0,\theta} + \beta_{1,\theta}x_1^s + \dots + \beta_{L,\theta}x_L^s) \\ q_\theta(x) &= \frac{\gamma_\theta^s}{\gamma_\theta^s - \gamma_\theta^d} (\alpha_{0,\theta} + \alpha_{1,\theta}x_1^d + \dots + \alpha_{K,\theta}x_K^d) - \frac{\gamma_\theta^d}{\gamma_\theta^s - \gamma_\theta^d} (\beta_{0,\theta} + \beta_{1,\theta}x_1^s + \dots + \beta_{L,\theta}x_L^s). \end{aligned}$$

Note that this is a generalization of the standard linear simultaneous equations model, which worries about only endogeneity from the intercepts, $\alpha_{0,\theta}$ and $\beta_{0,\theta}$. Here we can see the problem with simultaneity bias in the structural supply and demand equations: equilibrium price and quantity are functions of all the random parameters indexed by θ .

Take a given set of reduced form parameters π_θ , which are the composite coefficients in the above reduced forms. The corresponding structural parameter θ is identified under the well-known rank and order conditions. This discussion is summarized in the following theorem, where \mathcal{G}_π is the class of finite mixtures over reduced-form parameters.

¹³We restrict attention to the case of linear supply and demand models because identification in nonlinear models is quite complex (Benkard and Berry, 2006; Matzkin, 2009).

Theorem 4.3. Consider a parameter space Θ for the supply and demand model, where each $\theta \in \Theta$ satisfies

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

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.

We omit the proof because the previous discussion, an appeal to Theorem 4.1 to identify the reduced form and then an appeal to standard simultaneous identification results pointwise, is sufficient.¹⁴ The appeal to our reducibility condition is nested in the reference to Theorem 4.1, 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.

4.3 Literature Review for Nonparametric Regression

A literature focuses on the nonparametric identification of the distribution of random coefficients in 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 heterogeneous 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. This section shows how discrete choice models of demand are nonparametrically identified within our framework. Differentiated products are a special case of our framework; they are popular because they allow products to be distinguished by a parsimonious list of product

¹⁴The conditions that all demand and supply shifters never have zero coefficients is a bit strong; identification really requires that, for each θ , 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 $\gamma_{\theta}^d \neq -\gamma_{\theta}^s$ for all $\theta \in \Theta$. These assumptions ensure that there is a unique structural parameter θ that generates each reduced form π_{θ} .

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.

5.1 Discrete Choice Models Without Price Endogeneity

Let the utility from type θ purchasing product j be $u_\theta^j(x) + w_j$. A special case of this framework is when only the components of x corresponding to product j enter $u_\theta^j(x)$. The choice-specific scalar w_j is treated separately. One example is that w_j could be the price of good j . In this case, $u_\theta^j(x)$ is type θ 's reservation price for product j , and it would be more natural to write $u_\theta^j(x) - w_j$. However, w_j could be some non-price covariate or, with individual data, an interaction of a consumer and product characteristic. A consumer chooses j if $u_\theta^j(x) + w_j \geq u_\theta^k(x) + w_k \forall k \neq j$. A type $\theta \in \Theta$ indexes a J -tuple of such sub-utility functions $u_\theta(x) = (u_\theta^1(x), \dots, u_\theta^J(x))$.¹⁵ The goal is to identify the distribution $G(\theta)$ of utility functions for all J choices.

Implicit in the quasilinear representation of preferences $u_\theta^j(x) + w_j$ is the scale normalization that each type's coefficient on w_j is constrained to be 1. The normalization of the coefficient on w_j to be ± 1 is innocuous; choice rankings are preserved by dividing any type's utilities $u_\theta^j(x) + w_j$ by a positive constant. Thus if w admitted a type-specific coefficient $\alpha_\theta > 0$, then the type $\{u_\theta^j(x), \alpha_\theta\}$ would have the exact same preferences as the type $\left\{\frac{u_\theta^j(x)}{\alpha_\theta}, 1\right\}$. 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 $j = 0$ which has a utility of 0 for all types θ . We impose the tie-breaking rule that the good with the lowest index j is purchased in the event two or more goods have the same utility.

Assume that the product characteristics x varies over a set $X \subset \mathbb{R}^K$ 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 $(x, (w_1, \dots, w_J))$ varies over the product set $X \times \mathbb{R}^J$.

Theorem 5.1. *The distribution $G(\theta)$ over the type space Θ consisting of all vector analytic sub-utility functions u_θ is identified in \mathcal{G} .*

¹⁵The more typical empirical specification is when the functions $(u_\theta^1(x_1), \dots, u_\theta^J(x_J))$ are the same, up to a consumer-and-product-specific error term ϵ_θ^j , and where each product j has its own characteristics x_j . This is a special case of our framework.

Proof. Our strategy for proving identification works in two stages. Holding fixed x , we use variation in the regressors (w_1, \dots, w_J) to identify the joint distribution of values $u_{\tilde{\theta}} = (u_{\tilde{\theta}}^1, \dots, u_{\tilde{\theta}}^J) = (u_{\tilde{\theta}}^1(x), \dots, u_{\tilde{\theta}}^J(x)) \in \mathbb{R}^J$, for some $\theta \in \Theta$, in the population for each such $x \in X$. We use the notation $\tilde{\theta}$ because, conditional on x , two types θ_1 and θ_2 could have the same $u_{\tilde{\theta}}$. In the second stage, after the distribution of $u_{\tilde{\theta}}$ has been recovered for each menu of product characteristics $x \in X$, the resulting joint distribution $F(u_{\theta} | x)$ can be decomposed using the nonparametric regression model (Theorem 4.1) to recover the distribution over types $\theta \in \Theta$.

First, fix x , which gives the I -set for the outside good of

$$I_{0,w}^{\tilde{T}} = \left\{ \tilde{\theta} \in \tilde{T} \mid 0 > u_{\tilde{\theta}}^k + w_k \forall k = 1, \dots, J \right\},$$

where \tilde{T} is an arbitrary, finite subset of types with distinct $u_{\tilde{\theta}}$'s. Without loss of generality, we can focus on the case $\tilde{T} = I_{0,w}^{\tilde{T}} \cup \tilde{T}_3$ and $I_{0,w}^{\tilde{T}} = \{\tilde{\theta}_1, \tilde{\theta}_2\}$. We can always ensure a non-empty set by decreasing $w \in \mathbb{R}^J$. Types $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are different, so $u_{\tilde{\theta}_1}^j > u_{\tilde{\theta}_2}^j$ for some product j . Then we can raise w_j so that $u_{\tilde{\theta}_2}^j + w_j \geq 0$ but $u_{\tilde{\theta}_1}^j + w_j < 0$. Hence type $\tilde{\theta}_1$ drops out of the I -set but $\tilde{\theta}_2$ remains. Raising w_j will not cause any type $\tilde{\theta}_3 \in \tilde{T}_3$ to substitute to the outside good because of the key monotonicity assumption. We have reducibility of the submodel where x is fixed.

For any $x \in X$, we have identified the conditional, joint distribution of the subutilities, $G(u_{\tilde{\theta}}^1, \dots, u_{\tilde{\theta}}^J | x)$. As $u_{\tilde{\theta}}^j = u_{\theta}^j(x_j)$ for one or more θ 's, 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 subutility functions $u_{\theta} = (u_{\theta}^1(x), \dots, u_{\theta}^J(x))$. We can apply Theorem 4.1 to show that the distribution G over Θ is identified. Theorem 4.1 requires that each $u_{\theta} : X \rightarrow \mathbb{R}^J$ be vector analytic. \square

5.2 Price Endogeneity

We now consider the endogeneity problem that arises when θ is not independent of some elements of $(x, (w_1, \dots, w_J))$. Such endogeneity could arise if, for example, the choice set $(x, (w_1, \dots, w_J))$ that an agent faces is partly “designed” on the basis of information related to its type θ . A classic example of this source of endogeneity arises in a principal-agent relationship, in which the principal designs the menu of contracts $(x, (w_1, \dots, w_J))$ facing the agent using information that is correlated with the agent's type θ but that is not observable by the econometrician. The principal has incentives (i.e., screening) to use all information in contract design. Therefore, the endogenous choice of a menu of choices will induce a statistical endogeneity problem.

In this section, we show how to solve the endogeneity problem posed by endogenous product characteristics in multinomial choice by way of a triangular system of equations that builds on our work in section 4.1. Essentially, the system jointly models both the principal and the agent, and uses cross-sectional data on principal-agent relationships to achieve identification.

Let us denote the endogenous element of each product j as $p_j \in \mathbb{R}$. Hence the choice environment is described by

$$(p, x, w) = ((p_1, \dots, p_J), x, (w_1, \dots, w_J)).$$

We will let each p_j be the potentially endogenous attributes in the choice set, and thus the principal can design the menu (p_1, \dots, p_J) in some optimal fashion using information that is potentially correlated with the agent's type θ . The remaining choice characteristics $(x, (w_1, \dots, w_J))$ are exogenous and taken as given. Hence $p \not\perp \theta \mid (x, w)$, which is the source of the endogeneity problem (Pioner, 2008).

To handle the problem, we introduce the principal's pricing equation

$$p = g_\theta(x, w, z)$$

where $z = (z_1, \dots, z_J) \in \mathbb{R}^J$ is a vector of instruments that shifts the principal's pricing equation $f_\theta(x, w, z)$ in a locally full-rank way, i.e., for any $(x, w) \in X \times W$, $Dg_\theta(x, w, z)$ with respect to z has full rank J . Hence a type θ indexes a pair $\left(g_\theta, \left\{u_\theta^j(p, x)\right\}_{j=1}^J\right)$, where as usual we assume that all functions are real analytic. That is, a type indexes a principal agent relationship, where g_θ is potentially heterogeneous due to differing information sets or preferences among principals. Of course the joint distribution $G(\theta)$ over types allow the principal's pricing function g_θ to be stochastically dependent with the agent's preferences $\left\{u_\theta^j(p, x)\right\}_{j=1}^J$. The instruments (z_1, \dots, z_J) are most naturally interpreted as the marginal costs of providing each good, although they could represent any observed characteristics of the principal, including observed dimensions of its information set. By requiring only that each principal follow a different real analytic function for each type, we are relatively flexible about how principals make decisions. While we require the existence of a structure modeling the decisions of principals, we are nonparametric on that structure.

Assume that (p, x, z) has support containing an open and connected set V of the appropriate dimension, and (p, x, z, w) has support $V \times \mathbb{R}^J$. Our main result is that the endogenous multinomial choice model is reducible and hence identifiable.

Theorem 5.2. *The endogenous multinomial choice model is identified with respect to \mathcal{G} .*

Proof. Like the triangular system regression model the outcome variables of this model are the pair (p, j) , e.g. the principal's choice of prices and the agent's choice of product. Let $T \subset \Theta$ be a finite set of types, and consider the I -set

$$I_{p,0,x,w,z}^T = \left\{ \theta \in \Theta \mid g_\theta(x, w, z) \leq p \ \& \ u_\theta^j \left(g_\theta^j(x, w, z), x \right) + w_j < 0 \ \forall j \in \{1, \dots, J\} \right\},$$

where good 0 is the outside option that has a normalized utility of 0.

Following a similar logic to Theorem 4.2, we can adjust (p, x, w, z) so that $I_{p,0,x',w',z'}^T = I_{p,0,x,w,z}^T$

for all (x', w', z') in an open set U containing (x, w, z) . In particular, raise p and w by a sufficiently small amount so that all types in the I -set have g_θ strictly less than p and all types outside the I -set remain outside it.

Now there are two cases to consider. If $g_{\theta_1} \neq g_{\theta_2}$ for two types $\theta_1, \theta_2 \in T$, then we can replicate the first part of the proof for Theorem 4.2 to obtain reducibility via lowering p .

The second case to consider is when $g_{\theta_1} = g_{\theta_2}$ but $u_{\theta_1}^k \neq u_{\theta_2}^k$ for some inside good k . Then following the second part of the proof for Theorem 4.2, we can apply the open mapping theorem to find a $(x', w', z') \in U$ such that $u_{\theta_1}^k(f^k(x', w', z'), x'_k) \neq u_{\theta_2}^k(f^k(x', w', z'), x'_k)$. Now we simply raise w'_k so that one type switches to good k but the other type remains with the outside good. \square

We require only variation in the instruments z in a (potentially small) open set and do not require monotonicity assumptions about how the instrument affects the menu of choices through p_j or about how utility depends on p_j .

5.3 Support Conditions and the Pure Characteristics Demand Model

The previous theorems required large support on $w = (w_1, \dots, w_J)$ in order to induce any type to substitute away from the outside good. Otherwise, we could never distinguish between a type with a \$1 billion dislike of a product from a type with a \$1.1 billion dislike of the product. Notationally, we allow each type to have a separate subutility function $u_\theta^j(x)$ for each choice j , which subsumes the additive, i.i.d. error ϵ_θ^j common in the empirical literature.

Here we show by example that reducibility does not inherently rely on large-support assumptions. We work with the pure characteristics demand model (Bajari and Benkard, 2005; Berry and Pakes, 2007), where now $x = (x_1, \dots, x_J)$ and $u_\theta^j(x) = u_\theta(x_j)$ for all $j = 1, \dots, J$. In other words, the function $u_\theta(x_j)$ is common across choices up to the fact that each choice has different characteristics, x_j . Assume that $x = \{x_1, \dots, x_J\}$ lies in $X_1 \times \dots \times X_J$, where $X_1 = \dots = X_J$. Assume that $w = (w_1, \dots, w_J)$ lies in the product space $W_1 \times \dots \times W_J$, where $W_j = (0, \epsilon)$ for $\epsilon > 0$. Further, assume that for all $\theta \in \Theta$, there is a value x^θ such that $u_\theta(x^\theta) \in (0, \epsilon)$, an open interval. By continuity, there will be many such x^θ 's. Because ϵ can be small, we will not need large-support assumptions on w for identification.

Theorem 5.3. *The pure characteristics, multinomial choice model is identified with respect to \mathcal{G} .*

Proof. Turning to Definition 3.2, we choose the set of I -sets \mathbb{I}^T to contain only I -sets indexed by $x = \{x_1, \dots, x_J\}$ where each x_j satisfies $u_\theta(x_j) \in (0, \epsilon)$. The I -set with respect to choice 1 (for the case of exogenous characteristics) is

$$I_{1,x,w}^T = \{\theta \in \Theta \mid u_\theta(x_1) + w_1 \geq u_\theta(x_j) + w_j, 0 \forall j \in \{1, \dots, J\}, w_1 > 0\}.$$

Without loss of generality, we can restrict attention to $T = I_{1,x,w}^T \cup T_3$ and $I_{1,x,w}^T = \{\theta_1, \theta_2\}$, for finite $T \subseteq \Theta$. As in previous proofs, we can raise w_1 to w'_1 (to create a new vector w') so that there exists an open set $U \subseteq X$ where $I_{1,\bar{x},w'}^T = I_{1,\bar{x},w}^T$ for all $\bar{x} \in U$.

Let $v_\theta(a_1, a_2) = u_\theta(a_1) - u_\theta(a_2)$. As $u_\theta(\cdot)$ is a real-analytic function, so is $v_\theta(a_1, a_2)$. Thus there exists $(x'_1, x'_2) \in U$, so that $x' = (x'_1, x'_2, x_3, \dots, x_J)$, where $v_{\theta_1}(a_1, a_2) \neq v_{\theta_2}(a_1, a_2)$. Say $v_{\theta_1}(a_1, a_2) > v_{\theta_2}(a_1, a_2)$. Raise w_2 to w''_2 to create the new vector $w'' = (w'_1, w''_2, w_3, \dots, w_J)$. At some such (x', w'') , type θ_2 will substitute to product 2, but θ_1 will not. Raising w_2 to w''_2 will not encourage any $\theta_3 \in T_3$ to substitute to product 1. Thus, $I_{1,x',w''}^T$ reduces $I_{1,x,w}^T$.

By continuity, we can always make the open set U smaller so that the required w''_2 to induce θ_2 to substitute to product 2 is less than ϵ . \square

5.4 Purchasing Multiple Products with Complementarities or Substitutes in Preferences

Gentzkow (2007) and Liu, Chintagunta and Zhu (2008) study choice situations where each discrete choice $j = 0, \dots, J$ indexes a bundle of composite choices. For example, a consumer can purchase cable television separately ($j = 1$), purchase an internet connection separately ($j = 2$), purchase both cable television and an internet connection together as a bundle ($j = 3$), or purchase nothing, the outside good ($j = 0$). The goal in this situation is to distinguish between explanations for observed joint purchase: are consumers observed to buy cable television and internet connection at the same time because those who watch lots of television also have a high preference for television, or is there some causal utility increase from consuming both television and internet together? The goal is to distinguish unobserved heterogeneity in preferences for products, which may be correlated across products, from true complementarities.

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

$$\Delta_\theta(x) \equiv u_\theta^3(x) - (u_\theta^1(x) + u_\theta^2(x)).$$

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

Theorem 5.1 says we can identify the joint distribution of heterogeneity, which means we can identify the distribution of complementarities as a function of the joint distribution $G(\theta)$, if prices w_j are bundle-specific. Thus, we need to observe different choice situations where the bundle is or is not aggressively priced relative to the singleton packages. This is the data scheme for Liu et al.: they observe different bundles of telecommunications services at different prices, across geographic

markets.

5.5 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).¹⁶

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. For multinomial choice, the most commonly-used empirical model with unobserved heterogeneity is the random coefficients logit model. Bajari, Fox, Kim and Ryan (2009) were the first to prove the identification of the random coefficients logit model with continuous characteristics. They use calculus to show that all of the moments of the random coefficients are identified. The proof relies on linearity, $u_\theta^j(x) = x_j' \beta_\theta$, 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. There are two main differences. First, they focus on identifying the distribution of utility values conditional on characteristics, $G(u^1, \dots, u^J | x)$, rather than the distribution of utility functions, $u_\theta^1(x), \dots, u_\theta^J(x)$. This prevents them from calculating the distribution of utility changes after a policy intervention. It also prevents out-of-sample extrapolation: changing x outside its support. Second, Berry and Haile adopt a different approach to endogeneity. They require both individual and aggregate or market-level data and assume that the endogeneity occurs only in variables (like price) that vary at the market but not individual levels. They use individual data to trace out utility realization within a market and variation across markets to address an endogeneity problem. One could replace their step where they trace out utility values with our Theorem 5.1. Unlike methods 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 distributions of utility. We show above that reducibility does not inherently rely on large-support assumptions, because it can be applied to the pure characteristics demand model.

Studying the special case of $J = 2$, Ichimura and Thompson (1998) use the Cramer and Wold

¹⁶There is some work on multinomial discrete choice models examining the nonparametric identification of the distribution of a choice-specific error ϵ_θ^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.

(1936) theorem for identification, which relies critically on a linear index functional form: $u_\theta^j(x) = x_j' \beta_\theta$. We use only the quasilinearity of $u_\theta^j(x) + w_j$ in w_j and the vector 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 constant. 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.

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. We model the choice of product through a selection equation, which is simply a multinomial choice over products. The entire selection model is then a vector-valued demand model, where the dependent variable is the vector (j, y^j) .

A slight generalization of the previous theorems on multinomial choice and nonparametric regression allow us to identify the distribution of types when a type θ indexes the continuous outcome for choice 1 and the utility functions for all J choices: $\left\{ y_\theta^1, \left\{ u_\theta^j(x) \right\}_{j=1}^J \right\}$, and when the utility of a discrete choice is $u_\theta^j(x) + w_j$, as in the previous section. To see this, consider showing reducibility. There are two cases. First, two types θ_1 and θ_2 may differ in y_θ^1 . In that case, apply the nonparametric regression theorem. Second, two types may differ in their utility functions for the discrete choices. In that case, apply the multinomial choice theorem. Thus, our previous arguments are sufficient to identify the marginal distribution of an outcome for a particular choice, and the joint distribution of the outcome with the multinomial choice utility functions. The marginal distribution of the outcome y_θ^1 has been the object of interest in much but not all of the selection literature. In what follows, the new modeling structure and theorems will be devoted to identifying the joint distribution of all outcomes for all J choices. Full identification of the joint distribution of outcomes will distinguish our identification approach from some others in the literature.

6.1 The Roy Model

First, we consider the standard Roy model. A consumer picks product j if he will consume the highest quantity of j , conditional on purchase. Equivalently, a worker chooses the sector that pays the highest wage. An agent picks j if $y_\theta^j \geq y_\theta^k$ for $k = 1, \dots, j-1$ and $y_\theta^j > y_\theta^k$ for $k = j+1, \dots, J$. We use the tie-breaking rule that choice $j \in J$ is picked only if it is strictly preferred to all choices $j+1, \dots, J$. The equation governing consumption conditional on purchase is $y_\theta^j = f_\theta^j(x) + \beta_\theta^j w_j$, which is specific to choice j . Here x is a vector of characteristics that can implicitly include choice-specific characteristics, w_j is a characteristic specific to choice j , $\beta_\theta^j > 0$ (monotonicity) for all $\theta \in \Theta$, and $f_\theta^j(x)$ is a real analytic function. A type θ indexes $\{(\beta_\theta^1, \dots, \beta_\theta^J), (f_\theta^1(\cdot), \dots, f_\theta^J(\cdot))\}$. We will identify the joint distribution of the return functions for all J choices using information on the vector-valued dependent variable (j, y^j) . Note that we explore the identification of the distribution of all unknowns in the model and do not announce a treatment parameter as the object of interest, such as the average treatment effect (ATE). As before, identification will rely on exogenous characteristic variation in x . The assumptions on characteristics x and w are the same as in section 5.1. Note that we do not allow for an outside good, although an outside good (unemployment, say) could easily be added.

Theorem 6.1. *The Roy model is identified with respect to \mathcal{G} .*

Proof. We use Theorem 3.2, on two-step identification. We know from the nonparametric regression Theorem 4.1 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 if $\alpha_{\tilde{\theta}}^j + \beta_{\tilde{\theta}}^j w_j \geq \alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k$ for $k = 1, \dots, j-1$ and $\alpha_{\tilde{\theta}}^j + \beta_{\tilde{\theta}}^j w_j > \alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k$ for $k = j+1, \dots, J$, and where the covariates are allowed to vary as $w = (w_1, \dots, w_J) \in \mathbb{R}^J$. 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 distinct types $\tilde{\theta}$. Let $\{1, y, w\}$ be a triplet that defines

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

with at least two elements. Such a set exists by increasing w_1 and y . Without loss of generality, we can restrict attention to $T = I_{1,y,w}^T \cup T_3$ and $I_{1,y,w}^T = \{\tilde{\theta}_1, \tilde{\theta}_2\}$.

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

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 $I_{j,y,w'}$ and

$\tilde{\theta}_3$ for $\tilde{\theta}_3 \in T_3$ remains out of the I -set, 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 $I_{1,y',w'} = \{\tilde{\theta}_1\}$.

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

$$\alpha_{\tilde{\theta}}^k + \beta_{\tilde{\theta}}^k w_k = \alpha_{\tilde{\theta}}^1 + \beta_{\tilde{\theta}}^1 w_1 \quad (4)$$

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 $I_{1,y',w'} = \{\tilde{\theta}_1\}$ and, by monotonicity, $\tilde{\theta}_3 \notin I_{1,y',w'}$ for $\tilde{\theta}_3 \in T_3$. 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 (4), with w'_1 replacing w_1 , cannot involve $w_k^{\tilde{\theta}_1} = w_k^{\tilde{\theta}_2}$. Say $w_k^{\tilde{\theta}_1} > w_k^{\tilde{\theta}_2}$. 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

6.2 The Generalized Roy Model: Covariates in the Selection Equation

The generalized Roy model is a selection model that allows additional covariates to enter the selection equation. Let the utility for choice j be $f_{\theta}^j(x) + v_{\theta}^j(z) + \beta_{\theta}^j w_j$ and let the continuous outcome for choice j be $y_{\theta}^j = f_{\theta}^j(x)$. We require $f_{\theta}^j(x)$ and $v_{\theta}^j(z)$ to be real analytic functions. In our setup, the characteristics x affect only the continuous outcome while the characteristics z and the choice-specific scalar w_j affect only the discrete-choice utility. For simplicity we do not cover the case where some of the elements of x and z overlap, although that is a potential extension. We require the monotonicity assumption $\beta_{\theta}^j > 0$ for all $\theta \in \Theta$.¹⁷ Here θ indexes the collection of parameters and functions $\left\{ \left\{ \beta_{\theta}^j \right\}_{j=1}^J, \left\{ f_{\theta}^j(x) \right\}_{j=1}^J, \left\{ v_{\theta}^j(z) \right\}_{j=1}^J \right\}$. We identify the joint distribution $G(\theta)$ of the parameters indexed by θ . To complete the model, say an agent picks j if $f_{\theta}^j(x) + v_{\theta}^j(z) + \beta_{\theta}^j w_j \geq f_{\theta}^k(x) + v_{\theta}^k(z) + \beta_{\theta}^k w_k$ for $k = 1, \dots, j-1$ and $f_{\theta}^j(x) + v_{\theta}^j(z) + \beta_{\theta}^j w_j > f_{\theta}^k(x) + v_{\theta}^k(z) + \beta_{\theta}^k w_k$ for $k = j+1, \dots, J$. Let the assumptions on the supports of x , z and $w = (w_1, \dots, w_J)$ from section 5.2 hold.

A natural interpretation is that $f_{\theta}^j(x)$ 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 $f_{\theta}^j(x)$ would be the expected revenue he would earn in format j . One could study the

¹⁷There are two monotonicity assumptions: $\beta_{\theta}^j > 0$ and the discrete-choice utility is a positive function of the continuous outcome, $f_{\theta}^j(x)$. The normalization of the coefficient on $f_{\theta}^j(x)$ to be ± 1 is an innocuous scale normalization.

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 $f_\theta^1(x)$, or to sell without an agent, which would yield a return of $f_\theta^2(x)$.

Because $v_\theta^j(z)$ is part of a discrete choice utility function in a model without an outside good, we need an (innocuous) location normalization. Set $v_\theta^j(z) = 0$ for all z . Identification will require the Matzkin-like assumption $v_\theta^j(z^*) = a^* \forall j = 2, \dots, J$ and $\forall \theta$ for a normalizing pair z^* and a^* . It is not necessary for the normalizing pair to be known by the econometrician. We will discuss this assumption after the proof of the theorem.

Theorem 6.2. *The generalized Roy model is identified within the class \mathcal{G} .*

Proof. First we condition on x , the characteristics in the outcome equation. The problem will be to identify the distribution $G(\alpha_\theta^1, \dots, \alpha_\theta^J | x)$, where $\alpha_\theta^j = f_\theta^j(x)$ for one or more $\theta \in \Theta$. Then we use Theorem 3.2, on two-step identification. We know from the nonparametric regression Theorem 4.1 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 $\tilde{\theta}$ indexes $\left\{ \left\{ \beta_{\tilde{\theta}}^j \right\}_{j=1}^J, \left\{ \alpha_{\tilde{\theta}}^j \right\}_{j=1}^J, \left\{ v_{\tilde{\theta}}^j(z) \right\}_{j=1}^J \right\}$.

Consider a finite set T of distinct types $\tilde{\theta}$ and the I -set

$$I_{1,y,z,w}^T = \left\{ \tilde{\theta} \in T \mid [\alpha_{\tilde{\theta}}^1 + 0 + \beta_{\tilde{\theta}}^1 w_1 \geq \alpha_{\tilde{\theta}}^k + v_{\tilde{\theta}}^k(z) + \beta_{\tilde{\theta}}^k w_k \forall k = 2, \dots, J] \ \& \ [\alpha_{\tilde{\theta}}^1 \leq y] \right\}.$$

Such a set exists by increasing w_1 and y . Without loss of generality, we can restrict attention to $T = I_{1,y,z,w}^T \cup T_3$ and $I_{1,y,z,w}^T = \left\{ \tilde{\theta}_1, \tilde{\theta}_2 \right\}$.

The first case is when $\alpha_{\tilde{\theta}_1}^1 \neq \alpha_{\tilde{\theta}_2}^1$. We can just set $y' = \min \left\{ \alpha_{\tilde{\theta}_1}^1, \alpha_{\tilde{\theta}_2}^1 \right\}$ and the new I -set $I_{1,y',z,w}^T$ will reduce the old $I_{1,y,z,w}^T$.

Next, for some choice k , let $w_k^{\tilde{\theta}}(z, w_1)$ solve

$$\alpha_{\tilde{\theta}}^1 + 0 + \beta_{\tilde{\theta}}^1 w_1 = \alpha_{\tilde{\theta}}^k + v_{\tilde{\theta}}^k(z) + \beta_{\tilde{\theta}}^k w_k^{\tilde{\theta}}(z, w_1),$$

so that it is

$$w_k^{\tilde{\theta}}(z, w_1) = \frac{1}{\beta_{\tilde{\theta}}^k} \left[\alpha_{\tilde{\theta}}^1 + 0 + \beta_{\tilde{\theta}}^1 w_1 - (\alpha_{\tilde{\theta}}^k + v_{\tilde{\theta}}^k(z)) \right].$$

Division by $\beta_{\tilde{\theta}}^k$ is valid because $\beta_{\tilde{\theta}}^k > 0$ by the discrete choice monotonicity assumption. Each $w_k^{\tilde{\theta}}(z, w_1)$ is a real analytic function because all of its constituent terms are constants or real analytic functions. Form an open set $U \subset W \times Z$, where Z is the space of z , so that for all $(\hat{w}_1, \hat{z}) \in U$, $I_{1,y,z,w}^T = I_{1,y,\hat{z},\hat{w}_1}^T$. If $w_k^{\tilde{\theta}_1}(\cdot) \neq w_k^{\tilde{\theta}_2}(\cdot)$ as functions, then there exists $(w'_1, z') \in Z$ where $w_k^{\tilde{\theta}_1}(w'_1, z') \neq w_k^{\tilde{\theta}_2}(w'_1, z')$. Say $w_k^{\tilde{\theta}_1}(w'_1, z') > w_k^{\tilde{\theta}_2}(w'_1, z')$. Let w' be w with w'_1 in the first position and $w_k^{\tilde{\theta}_2}(w'_1, z')$ in the k th position. Then $\tilde{\theta}_2$ switches to product k at (w', z') and so

$I_{1,y,z',w'}^T = \{\tilde{\theta}_1\}$ reduces $I_{1,y,z,w}^T$.

The challenge is to show $w_k^{\tilde{\theta}_1}(\cdot) \neq w_k^{\tilde{\theta}_2}(\cdot)$ as functions for some choice k , i.e. there exists a least one point (w_1'', z'') where $w_k^{\tilde{\theta}_1}(w_1'', z'') \neq w_k^{\tilde{\theta}_2}(w_1'', z'')$ for some choice k . The function $w_k^{\tilde{\theta}}(z, w_1)$ can be rewritten as

$$w_k^{\tilde{\theta}}(z, w_1) = A_{\tilde{\theta}} - \frac{v_{\tilde{\theta}}^k(z)}{\beta_{\tilde{\theta}}^k} + \frac{\beta_{\tilde{\theta}}^1}{\beta_{\tilde{\theta}}^k} w_1,$$

where $A_{\tilde{\theta}} = (\alpha_{\tilde{\theta}}^1 - \alpha_{\tilde{\theta}}^k) (\beta_{\tilde{\theta}}^k)^{-1}$ is a constant in z and w_1 . Recall $\beta_{\tilde{\theta}}^k > 0$ and $\beta_{\tilde{\theta}}^1 > 0$.

First, consider the case where $v_{\tilde{\theta}_1}^k \neq v_{\tilde{\theta}_2}^k$ for some k . Now $(v_{\tilde{\theta}_1}^k(z) - v_{\tilde{\theta}_2}^k(z)) \neq B_{\tilde{\theta}}$, where $B_{\tilde{\theta}}$ is a constant function, because of the normalization that $v_{\tilde{\theta}}^j(z^*) = a^*$ for all $\tilde{\theta}$. By inspection, the partial derivative of $w_k^{\tilde{\theta}_1}(z, w_1)$ in some element of z will not equal the partial derivative of $w_k^{\tilde{\theta}_2}(z, w_1)$ in the same element of z , so the functions are different.

Second, consider the case where $v_{\tilde{\theta}_1}^k = v_{\tilde{\theta}_2}^k$ for all choices k and $\beta_{\tilde{\theta}_1}^j \neq \beta_{\tilde{\theta}_2}^j$ for some choice j . If $j = k \geq 2$, then the partial derivative of $w_k^{\tilde{\theta}_1}(z, w_1)$ in any element of z will not equal the partial derivative of $w_k^{\tilde{\theta}_2}(z, w_1)$ in the same element of z , so the functions are different. If $j = 1$, then $\frac{\partial w_k^{\tilde{\theta}}(z, w_1)}{\partial w_1} = \frac{\beta_{\tilde{\theta}}^1}{\beta_{\tilde{\theta}}^k}$ differs between $\tilde{\theta}_1$ and $\tilde{\theta}_2$ unless $\frac{\beta_{\tilde{\theta}_1}^1}{\beta_{\tilde{\theta}_1}^k} = \frac{\beta_{\tilde{\theta}_2}^1}{\beta_{\tilde{\theta}_2}^k}$, but the latter possibility can only happen when $\beta_{\tilde{\theta}_1}^j \neq \beta_{\tilde{\theta}_2}^j$ also for $j = k \geq 2$. We would then apply the $j = k \geq 2$ case in this paragraph.

We studied the case $\alpha_{\tilde{\theta}_1}^1 \neq \alpha_{\tilde{\theta}_2}^1$, so the only remaining case is when all random elements are the same between $\tilde{\theta}_1$ and $\tilde{\theta}_2$ except for $\alpha_{\tilde{\theta}}^k$ for at least one choice $k \geq 2$. Clearly $A_{\tilde{\theta}_1} \neq A_{\tilde{\theta}_2}$ in this case. As all other elements are the same, $w_k^{\tilde{\theta}_1}(\cdot) \neq w_k^{\tilde{\theta}_2}(\cdot)$ as functions. \square

The proof uses two types of dependent-variable variation. First, for types who differ in their continuous outcomes for choice 1, we need examine only variation in the continuous outcome for choice 1. Second, for two types who differ in some other component, we must induce one of them to switch to another product before the other. This second proof approach prevents us using a utility function like $\alpha_{\tilde{\theta}}^j + v_{\tilde{\theta}}^j(z) + \beta_{\tilde{\theta}}^j w_j$ where $v_{\tilde{\theta}}^j(z) = \bar{v}_{\tilde{\theta}}^j(z) + \epsilon_{\tilde{\theta}}^j$ and $\epsilon_{\tilde{\theta}}^j$ is an additive error term in the discrete choice utility. The Matzkin-like normalization rules this case out. Otherwise, two types with $\alpha_{\tilde{\theta}_1}^1 + \epsilon_{\tilde{\theta}_1}^1 - (\alpha_{\tilde{\theta}_1}^k + \epsilon_{\tilde{\theta}_1}^k) = \alpha_{\tilde{\theta}_2}^1 + \epsilon_{\tilde{\theta}_2}^1 - (\alpha_{\tilde{\theta}_2}^k + \epsilon_{\tilde{\theta}_2}^k)$ will switch from product 1 to product k at the same point, if all other random elements are identical between the two types. The proof would require using variation in y_k to distinguish why some type switched: a high continuous outcome or a high discrete-choice utility? However, the condition in reducibility, Definition 3.2, that the set $I_{y',x'}^T$ that reduces $I_{y,x}^T$ be a strict subset rules this out: we cannot examine continuous outcomes for choice k without allowing types who were all along buying product k into the I -set.

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. We use reducibility to show identification of both the outcome (usage) and selection (product choice) equations.

To some degree, our approach to selection harkens 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 (Heckman, Urzua and Vytlacil, 2006). 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 and 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 the generalized Roy 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 .¹⁸ 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. 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 of the the generalized Roy model, 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

¹⁸Chamberlain (1986, page 205) defines the term identification at infinity. In generic notation, let $Q(x)$ be a function of data x and let β be a parameter to be identified. The parameter β is identified at infinity if $Q(x) \neq \beta$ for $x < \infty$ but $\lim_{x \rightarrow \infty} Q(x) = \beta$.

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 are less sensitive to the tail frequencies of $y_1 - y_2$. The traditional literature does not focus on medians because median treatment effects require knowledge of the joint distribution of the outcomes, 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 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 identifying the joint distribution of outcomes allows us to compute these measures.

6.3.2 Allowing Random Coefficients and Functions 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.¹⁹

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 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 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 require one choice-specific variable w_j , in either the outcome or selection equations, to have support that is either strictly negative or strictly positive. Even with a common sign, allowing for random coefficients on this variable means the magnitude of the effect of substituting towards or away from various options

¹⁹See 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), Bayer, Khan and Timmins (2008) and many others.

can vary quite a bit across the population. We impose no monotonicity restrictions on the $v_{\theta}^j(z)$ functions. We also impose no conditions on the joint dependence of each $v_{\theta}^j(z)$ function with the J outcome equations, $f_{\theta}^j(x)$. Say z includes the price of all J products. Then we allow those who are more sensitive to price to be more sensitive to the product characteristics in x in their continuous outcomes.

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 reducibility 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 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.²⁰ Of course, our object of interest is an entire distribution function and not a scalar parameter, such as an average treatment effect. However, the analysis in Bajari, Fox, Kim and Ryan (2009) for a particular estimator suggests that estimating the distribution as a mixture and then computing a scalar

²⁰For special cases of selection and endogeneity, Hong and Tamer (2003), Vytlačil and Yildiz (2007), and Shaikh and Vytlačil (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.

function of the distribution (a mean, say) can result in an estimator for the scalar parameter that is \sqrt{n} -consistent and asymptotically normal. These results are obtained by placing the estimator in the sieve framework of Chen and Pouzo (2008).

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

$$E \left[y_2 \mid \{x_j, z_j\}_{j=1}^2 \right] - E \left[y_1 \mid \{x_j, z_j\}_{j=1}^2 \right].$$

There is no conditioning on the product choice $D \in \{1, 2\}$ in the definition of the average treatment effect. Their bounds point identify the average treatment effect under identification at infinity. If the exogenous data do not allow the researcher to observe a situation where $\Pr(D = 2) = 1$, Heckman and Vytlacil prove their bounds are sharp: a model can be constructed that is 1) consistent with the data on $\{y_D, D, \{x_j, z_j\}_{j=1}^2\}$ 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 Vytlacil 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_D and D . Our assumptions and those of Heckman and Vytlacil are non-nested.

7 Estimation: A Monte Carlo 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 mixtures identification strategy naturally suggests the linear regression mixtures estimator of Bajari, Fox, Kim and Ryan (2009). The Monte Carlo experiment demonstrates 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. Further, the Monte Carlo experiment is an example that suggests that selection models are identified under more general conditions than this paper uses. The example

uses random coefficients that have continuous support and the discrete choice model does not have a monotonicity assumption.

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 generated 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. 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 the regression of $y_{i,2}$ on $x_{2,2}$ gives -0.50, understating the mean by 0.50. Finally, 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,j_i} on the observed $x_{i,j_i,2}$ 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 $\hat{\pi}^r$

on each of the R normal components, the final estimator for $g(\beta_1, \beta_2, \beta_3)$ is

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

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

$$1[j_i^* = j^s, y_{i,j} \leq y^s] = \sum_{r=1}^R \pi^r \cdot \int_{\beta_1} \int_{\beta_2} \int_{\beta_3} H(\beta, x_i, j^s, y^s) \phi(\beta_1 - \beta_1^r) \phi(\beta_1 - \beta_2^r) \phi(\beta_1 - \beta_3^r) d\beta_1 d\beta_2 d\beta_3 \quad (5)$$

where the expression

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

represents whether an agent with preferences β would both select choice j^s and have a continuous outcome $y_j \leq y^s$, with the covariates of the statistical observation i . We choose $S = 62$, which corresponds to the 2 binary choices times 31 evenly spaced y^s points, from -10.5 to 12.7. The three-dimensional numerical integral in (5) 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 $N \cdot S$ regression observations in a linear regression where the only unknown parameters are the R π^r 's. As the unknown parameters π^r enter (5) linearly, the estimator is linear regression subject to the inequality constraints $\pi^r \geq 0 \forall r = 1, \dots, R$ and $\sum_{r=1}^R \pi^r = 1$. 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

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 small 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 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.

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.²¹ 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, (β_2, β_3) , because the researcher only observes data on only one of these outcomes at a time. Here we examine whether our mixtures identification strategy can identify the bivariate distribution of outcomes. Figure 5 plots the joint density of the two outcome equations. 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.

²¹Other experiments suggest $R = 75$ may be too low; we might need $R = 500$ for a near perfect approximation to the joint density.

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. We introduce a property of economic models, known as reducibility, that is a sufficient condition for identification of the distribution of heterogeneity.

We apply reducibility to three classes of models: discrete choice models, nonparametric regression 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 choice environment x to enter a nonparametric function $f_\theta(x)$ that is known only to be real analytic. We are nonparametric on the distribution $G(\theta)$ of the functions $f_\theta(x)$. 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 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 do not rely on the assumptions of 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 utility functions for all choices, 5) we are nonparametric on the subutility function $u_\theta^j(x_j)$ for choice j , 6) we allow for endogenous characteristics such as prices, 7) we show how to analyze multiple purchases when some goods can be complements, and 8) we show that we do not need large support if demand is given by the pure characteristics model.

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, and 4) our argument generalizes easily to the case of multinomial choice in the selection equation.

Our mixtures identification strategy, while not constructive, does naturally suggest the linear regression estimator of Bajari, Fox, Kim and Ryan (2009). We conduct a Monte Carlo study 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.

A Identification With Positive Probability

Consider the mixtures model (1). To show the consistency of a nonparametric mixtures estimator, one typically needs a stronger definition of identification than is used in the statistics literature following Teicher (1963). For two distributions G^0 and G^1 , one needs that there exists a set $X^* \subseteq X$ with *positive probability* such that for all $x \in X^*$, $F_{G^0}(y | x) \neq F_{G^1}(y | x)$ for some fixed y . We can this strong definition of identification “identification with positive probability.”

As we now show, the existence of such a set positive measure X^* follows readily from the existence of a single $(y, x) \in Y \times X$ for which $F_{G^0}(y | x) \neq F_{G^1}(y | x)$, as ensured by reducibility. In particular, we show that from the existence of such an experiment (y, x) , we can find a small open ball X^* about x .

Lemma A.1. *Identification implies identification with positive probability if for any finite set of types $T \subset \Theta$, and for any $I_{y,x}^T$, there exists (y', x') such that $I_{y,x}^T = I_{y',x'}^T$ and for some small neighborhood $X^* \subset X$ containing x' , $z \in X^*$ implies $\exists y_z$ such that $I_{y_z,z}^T = I_{y',x'}^T$*

Proof. We can always define G^0 and G^1 to assign probabilities to the same set of finite types $T = \{\theta_1, \dots, \theta_n\} \subset \Theta$ by simply taking the union of their supports, $T = T^0 \cup T^1$, and adding zero probability masses where necessary. Thus G^0 and G^1 can each be represented by, respectively, points of the form p_θ^0 and p_θ^1 . Let (y, x) be the experiment that distinguishes G^0 and G^1 . Then we have

$$\sum_{\theta \in I_{y,x}^T} p_\theta^0 \neq \sum_{\theta \in I_{y,x}^T} p_\theta^1.$$

Also, for each $z \in X^*$ we have

$$\sum_{\theta \in I_{y_z,z}^T} p_\theta^0 \neq \sum_{\theta \in I_{y_z,z}^T} p_\theta^1.$$

□

We next consider two examples of applying this lemma.

Theorem A.1. *The nonparametric regression model is identified with positive probability in the class \mathcal{G} .*

Proof. Consider $I_{y,x}^T = \{\theta \in T \mid f_\theta(x) \leq y\}$. Raise y by a sufficiently small amount so no new types enter the I -set, yielding y' . Now each type in the I -set has a response strictly less than

y' , and each type not in the I -set has response strictly greater than y . As each type's $f_\theta(x)$ is continuous in $x \in X$, these inequalities are preserved for all $z \in X^*$ where X^* is a sufficiently small neighborhood containing x . Apply Lemma A.1. \square

Theorem A.2. *The multinomial choice model is identified with positive probability in the class \mathcal{G} .*

Proof. Consider the first of the two parts of the proof of Theorem 5.1, as the second part falls under the proof of Theorem A.2 above. The I -set for the finite set \tilde{T} is

$$I_{0,w}^{\tilde{T}} = \left\{ \tilde{\theta} \in \tilde{T} \mid 0 > u_{\tilde{\theta}}^k + w_k \forall k = 1, \dots, J \right\}.$$

Raise each w_k a sufficiently small amount so that no types leave the I -set. By monotonicity, none will join the I -set. Then there is a W^* where $I_{0,w'}^{\tilde{T}} = I_{0,w}^{\tilde{T}}$ for $w' \in W^*$. Apply Lemma A.1. \square

B Reducibility When Distributions Admit a Density Function

We wish to show that an analog to the concept of reducibility can be extended to models where the distribution $G(\theta)$ is required to admit a continuous density $g(\theta)$ over Θ . This is a non-nested class to the class of multinomial mixtures that we study. While we do not believe the extension is essential for practical applications in economics, it is of some theoretical interest. As before, let the economic model be (\mathcal{M}, X) . For any $T \subset \Theta$, the I -set $I_{y,x}^T = \{\theta \in T \mid f_\theta(x) \leq y\}$ may no longer be a finite set of points. In certain well-behaved models, we may imagine $I_{y,x}^T$ to be a connected subset of T .

Let $\mathcal{A} = \{A_k \mid k \in \mathcal{K}\}$ be the class of all sets such that for each $k \in \mathcal{K}$, $A_k \subseteq \Theta$ is the union of disjoint, connected, open sets: $A_k = \bigcup_{i \in C_{A_k}} U_i$, where C_{A_k} is the index set for the disjoint sets in A_k . An economic model (\mathcal{M}, X) is **\mathcal{A} -reducible** if, for every $T \in \mathcal{A}$ ($T = A_k$ for some $k \in \mathcal{K}$) and $T \subseteq \Theta$, where $T = \bigcup_{i \in C_T} U_i$, there exists $(y, x) \in \mathbb{R}^m \times X$ and $i \in C_T$ for which $I_{y,x}^T \subseteq U_i$ for some open, connected $U_i \subset T$.

Theorem B.1. *If (\mathcal{M}, X) is \mathcal{A} -reducible, then it is identified within the class of distributions with continuous densities over Θ .*

Proof. Suppose that \mathcal{A} -reducibility holds but that the model is not identified. Then, there exist two continuous densities $g_0(\theta)$, the truth, and $g_1(\theta)$ that both give the same distribution $F(y \mid x) = F_0(y \mid x) = F_1(y \mid x)$ for the data. Let $\pi(\theta) = g_0(\theta) - g_1(\theta)$. The function $\pi(\theta)$ is continuous

because $g_0(\theta)$ and $g_1(\theta)$ are. Then

$$F_0(y|x) - F_1(y|x) = \int_{\Theta} \pi(\theta) 1[f_{\theta}(x) \leq y] d\theta = 0.$$

Define $\pi^+(\theta) = \pi(\theta) 1[\pi(\theta) \geq 0]$ and $\pi^-(\theta) = -\pi(\theta) 1[\pi(\theta) < 0]$, so that $\pi(\theta) = \pi^+(\theta) - \pi^-(\theta)$. Therefore,

$$\int_{\Theta} \pi^+(\theta) 1[f_{\theta}(x) \leq y] d\theta = \int_{\Theta} \pi^-(\theta) 1[f_{\theta}(x) \leq y] d\theta. \quad (6)$$

By the continuity of $\pi(\theta)$, $\pi^+(\theta)$ and $\pi^-(\theta)$ have disjoint and open supports. Let T be the union of these supports: a union of disjoint, connected, open sets in which either $\pi^+(\theta) > 0$ or $\pi^-(\theta) > 0$ on any one of these open sets. Therefore, T is in \mathcal{A} . By \mathcal{A} -reducibility, there exists $(y, x) \in \mathbb{R}^m \times X$ and $i \in C_T$ for which $I_{y,x}^T \subseteq U_i$ for some open, connected $U_i \subset T$. However, either

$$\int_{\Theta} \pi^+(\theta) 1[f_{\theta}(x) \leq y] d\theta = \int_{I_{y,x}^T} \pi^+(\theta) 1[f_{\theta}(x) \leq y] d\theta \neq 0$$

and the equivalent expression for $\pi^-(\theta)$ equals 0, or vice versa. This is a contradiction to (6), and so we have identification. \square

C Proof of Two-Step Identification, Theorem 3.2

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

$$G_c(z_1, z_2) = \Pr(\theta_1 \leq z_1, \alpha(c, \theta_2) \leq z_2) = G(\{\theta \in \Theta : \theta_1 \leq z_1, \alpha(c, \theta_2) \leq z_2\}). \quad (7)$$

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 3.1 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 (7). 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, $\theta_1^k < \tilde{\theta}_1^k$. By exploiting the properties of a partial order, 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 $F(y | x_1, x_2 = c)$ implied by G and G' via

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

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 $(f, X_1, \mathbb{R}^{N_1+N_2})$, there exists $x_1 \in X_1$ for which the distribution function $F(\cdot | x_1, x_2 = c) \neq F'(\cdot | x_1, x_2 = c)$. Hence we have identification.

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 and C. Lanier Benkard**, “Demand Estimation With Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach,” *The Journal of Political Economy*, 2005, 113 (6), 1239–1276.
- , **Jeremy T. Fox, Kyoo il Kim, and Stephen Ryan**, “Identification and Estimation of Random Utility Models,” March 2009. working paper.
- Barbe, Philippe**, “Statistical analysis of mixtures and the empirical probability measure,” *Acta Appl. Math.*, 1998, 50 (3), 253–340.

- Bayer, Patrick, Shakeeb Khan, and Christopher Timmins**, “Nonparametric Identification and Estimation in a Generalized Roy Model,” March 2008. working paper.
- Benkard, C.L. and S. Berry**, “On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on Brown (1983) and Roehrig (1988),” *Econometrica*, 2006, *74* (5), 1429–1440.
- Beran, R. and PW Millar**, “Minimum Distance Estimation in Random Coefficient Regression Models,” *The Annals of Statistics*, 1994, *22* (4), 1976–1992.
- Berry, Steven and Ariel Pakes**, “The Pure Characteristics Demand Model,” *International Economic Review*, 2007, *48* (4), 1193–1225. working paper.
- 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.
- Chen, Xiaohong and Damian Pouzo**, “Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals,” *Journal of Econometrics*, October 2008.
- 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. Vytlacil**, “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.
- Gentzkow, Matthew**, “Valuing New Goods in a Model with Complementarity: Online Newspapers,” *The American Economic Review*, 2007, *97* (3), 713–744.
- 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, James 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), 153–162.
- , “Varieties of Selection Bias,” *The American Economic Review*, 1990, *80* (2), 313–318.
- and **Burton S. Singer**, “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data,” *Econometrica*, 1984, *52* (2), 271–320.
- and **Edward J. Vytlačil**, “Local instrumental variables and latent variable models for identifying and bounding treatment effects,” *Proceedings of the National Academy of Sciences*, 1999.
- and **Salvador Navarro**, “Dynamic Discrete Choice and Dynamic Treatment Effects,” *Journal of Econometrics*, 2007.
- , **Sergio Urzua**, and **Edward J. Vytlačil**, “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, 2006, *88* (3), 389–432.
- Heckman, J.J. and B.E. Honore**, “The Empirical Content of the Roy Model,” *Econometrica*, 1990, *58* (5), 1121–1149.
- and **E.J. Vytlačil**, “Instrumental variables, selection models, and tight bounds on the average treatment effect,” *Econometric Evaluation of Labour Market Policies*, 2001.
- , **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.
- 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.
- Krantz, Steve G. and Harold R. Parks**, *A Primer on Real Analytic Functions*, second ed., Birkhäuser, 2002.
- 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.
- Liu, Hongju, Pradeep Chintagunta, and Ting Zhu**, “Complementarities and the Demand for Home Broadband Internet Services,” October 2008. working paper.
- 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.
- , “Identification in Nonparametric Simultaneous Equations,” *Econometrica*, 2009.
- McFadden, Daniel L. and Kenneth Train**, “Mixed MNL Models for Discrete Response,” *Journal of Applied Econometrics*, 2000, 15, 447–470.
- Pioner, Heleno**, “Semiparametric Identification of Multidimensional Screening Models,” 2008. working paper.
- 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, 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, 1 (1), 40–69.
- Vytlačil, 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. Vytlačil, eds., *Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics*, Vol. 21, Elsevier, 2007.
- Yakowitz, S.J. and J.D. Spragins**, “On the Identifiability of Finite Mixtures,” *The Annals of Mathematical Statistics*, 1968, 39 (1), 209–214.

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

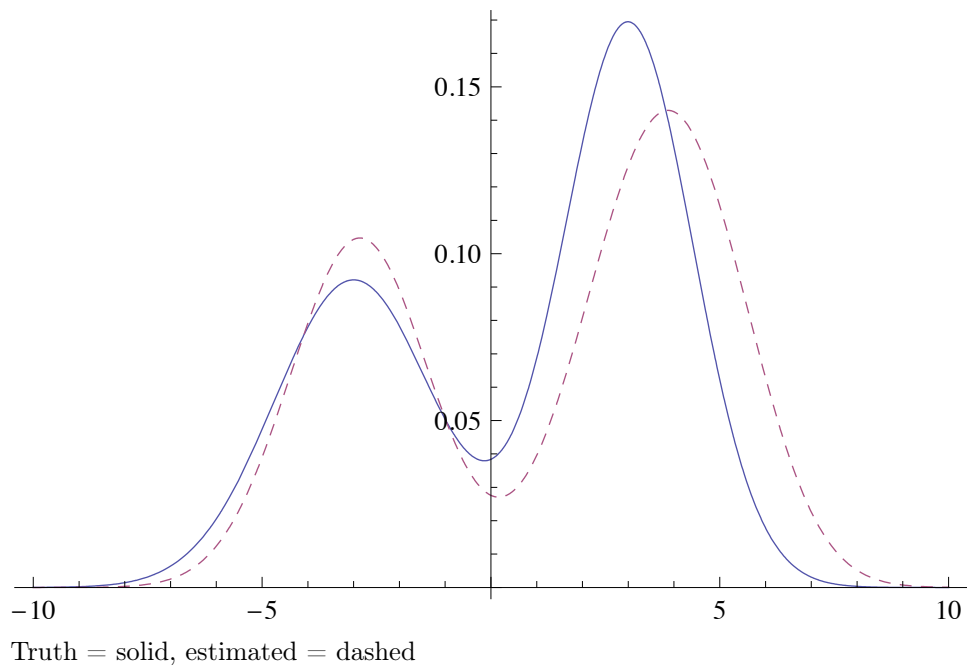


Figure 2: True and Estimated Marginal Density of the Intercept in the Outcome Equation For Choice 1, $\beta_{i,2}$

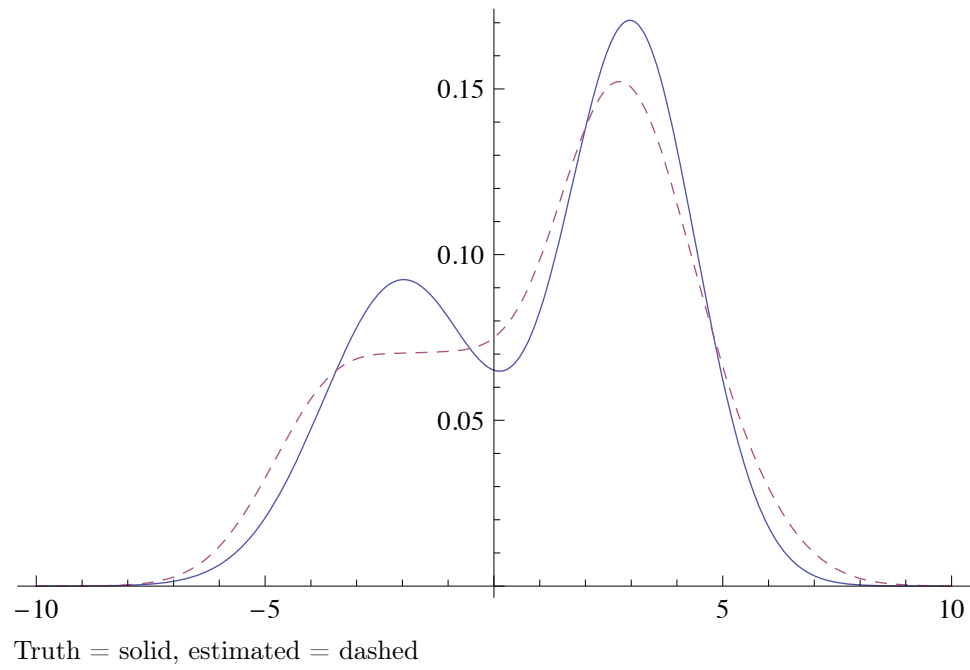


Figure 3: True and Estimated Marginal Density of the Intercept in the Outcome Equation For Choice 2, $\beta_{i,2}$

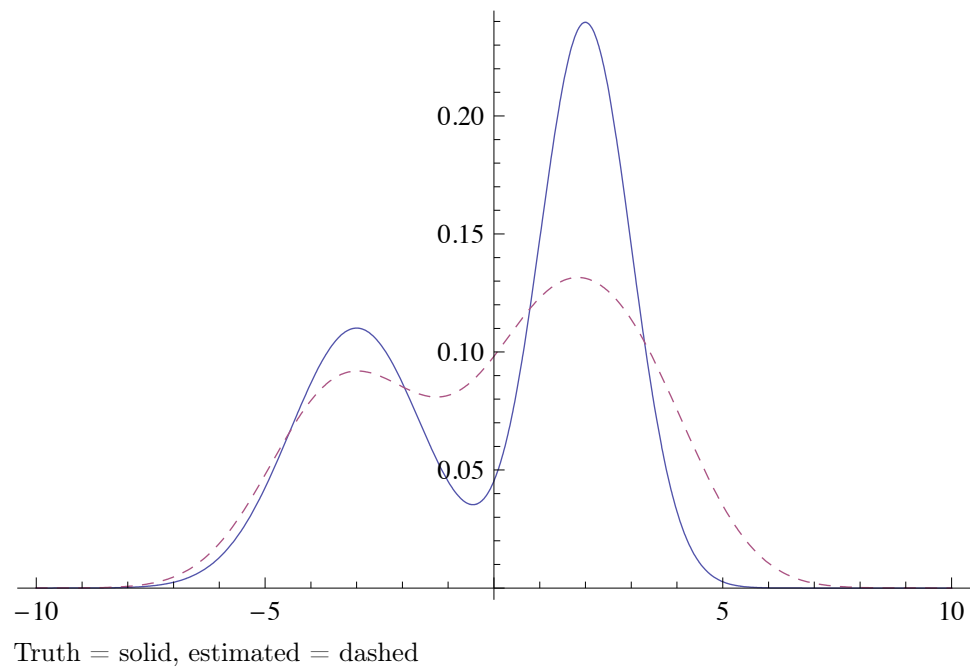


Figure 4: True and Estimated Joint Density of the Selection Slope and Intercept in the Outcome Equation For Choice 1, $\beta_{i,1}$ and $\beta_{i,2}$

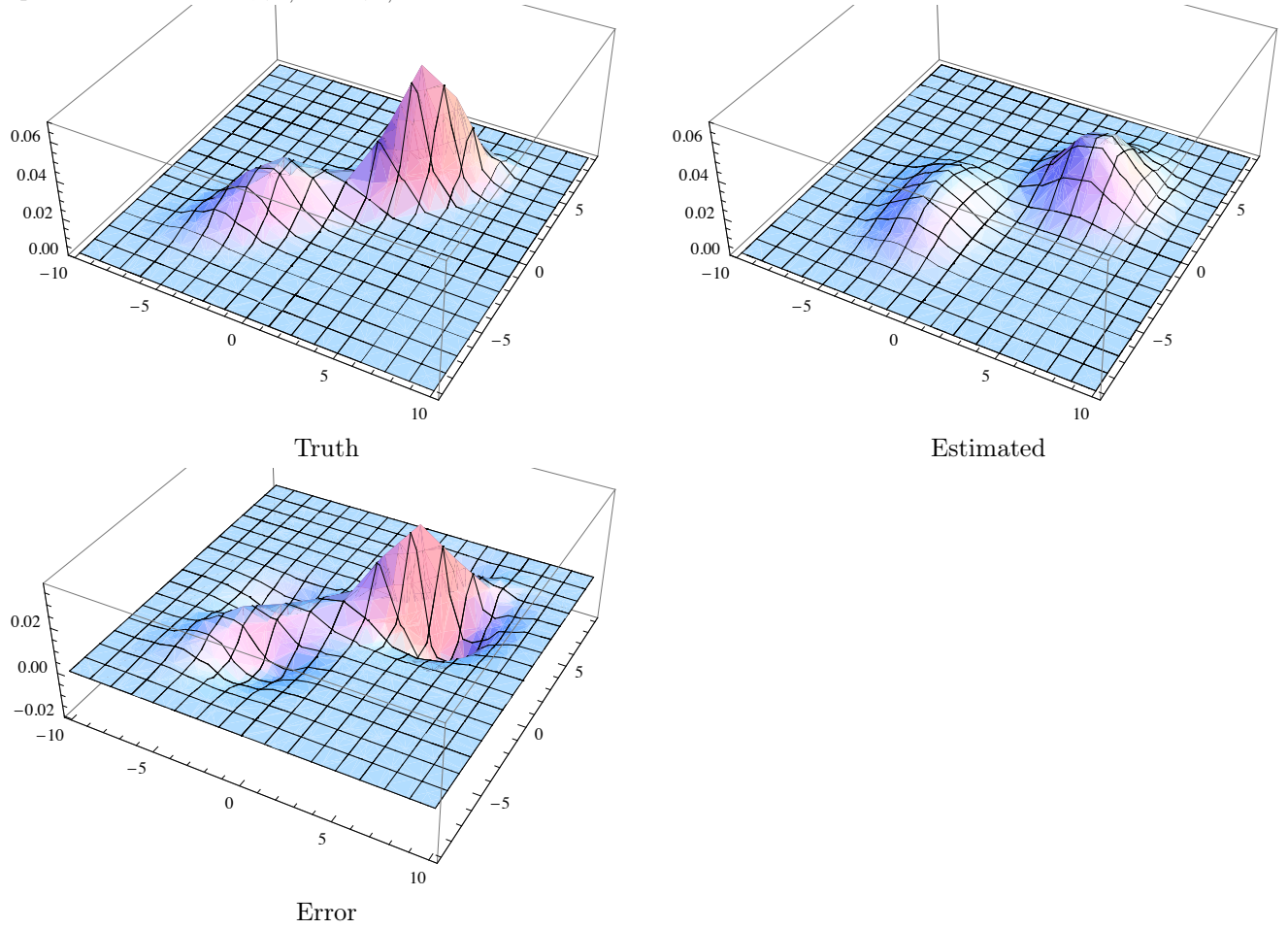


Figure 5: True and Estimated Joint Density of the Intercepts in the Outcome Equations For Both Choices 1 and 2, $\beta_{i,2}$ and $\beta_{i,3}$

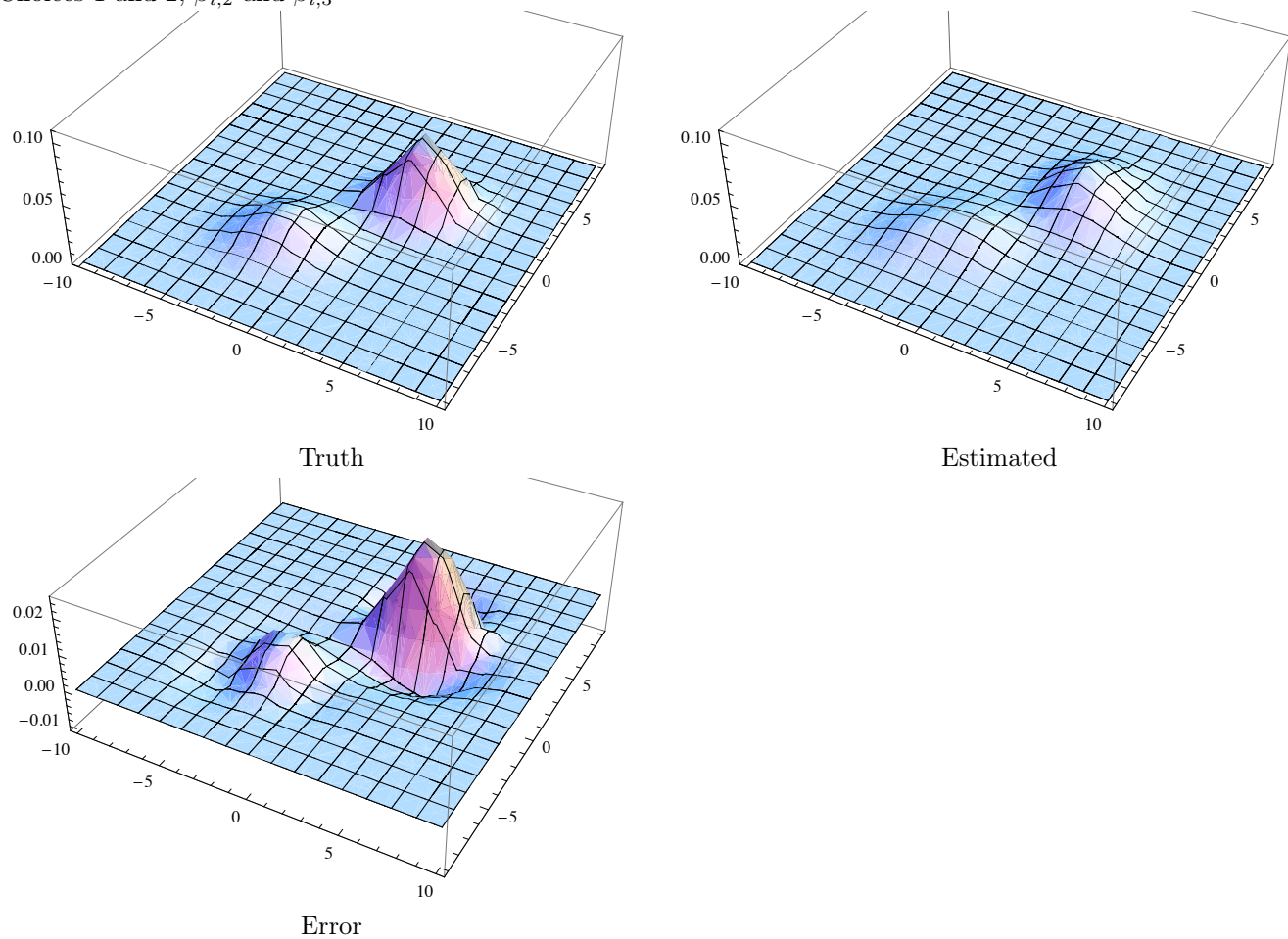


Table 1: True and Estimated Means of the Random Coefficients

| | 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 |

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.