NBER WORKING PAPER SERIES

IDENTIFYING HETEROGENEITY IN ECONOMIC CHOICE MODELS

Jeremy T. Fox
Amit Gandhi

Working Paper 15147
http://www.nber.org/papers/w15147

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2009

Thanks to Steven Durlauf, James Heckman, Salvador Navarro, Philip Reny, Azeem Shaikh, Susanne Schennach, Morten Sørensen, Harald Uhlig and Edward Vytlacil for helpful comments. Also thanks to seminar participants at Brown, the Brown / UCL Demand Conference, Caltech, CREST, Chicago, Cowles, EC2 Rome, LSE, Michigan, Michigan State, Minnesota, Northwestern, Rochester, the SED, Stanford, Toulouse, UCL, USC, Washington University in St. Louis, Wisconsin and Yale. Fox thanks the National Science Foundation, the Olin Foundation, and the Stigler Center for financial support. Thanks to Chenchuan Li for research assistance. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Identifying Heterogeneity in Economic Choice Models
Jeremy T. Fox and Amit Gandhi
NBER Working Paper No. 15147
July 2009
JEL No. C14,C25,L0

ABSTRACT

We show how to nonparametrically identify the distribution that characterizes heterogeneity among agents in a general class of structural choice models. We introduce an axiom that we term separability and prove that separability of a structural model ensures identification. The main strength of separability is that it makes verifying the identification of nonadditive models a tractable task because it is a condition that is stated directly in terms of the choice behavior of agents in the model. We use separability to prove several new results. We prove the identification of the distribution of random functions and marginal effects in a nonadditive regression model. We also identify the distribution of utility functions in the multinomial choice model. Finally, we extend 2SLS to have random functions in both the first and second stages. This instrumental variables strategy applies equally to multinomial choice models with endogeneity.

Jeremy T. Fox
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
fox@uchicago.edu

Amit Gandhi
University of Wisconsin
1180 Observatory Drive
Madison, WI 53706-1393
agandhi@ssc.wisc.edu

1 Introduction

Heterogeneity among decision makers, be they firms or consumers, is a critical feature of economic life that is important for the study of many policy problems. For example, some consumers may value a product characteristic more than others, so that the consumers with higher values might have less elastic demands. Likewise, more productive firms may have an incentive to adopt a new technology sooner, so that the returns of early adopters exceed the returns of late adopters.

Increasingly, researchers in industrial organization have begun to analyze the consumption choices of individual consumers and the production decisions of individual firms. In so far as the underlying heterogeneity in tastes across consumers and heterogeneity in technologies across firms is unobserved to the econometrician, the structural error term in the choice model enters in a generally non-additive way. Furthermore, the form of heterogeneity among agents is not easily indexed by a finite vector, but rather indexed more naturally by a function, i.e., a utility function that characterizes a consumer or a production function that characterizes a firm.

This paper presents a general mathematical approach to establishing the identification of the distribution of heterogeneity in choice models in which agents are indexed by functions that capture their tastes, technologies, etc. Our identification results are nonparametric in two respects: we do not impose parametric assumptions on the functions that characterize individual agents or on the distribution of heterogeneity. Nonparametric and flexibly parametric estimators have been proposed for estimating the distribution of heterogeneity in structural models. However, less work has been done showing the identification of such models. Without showing identification, a full proof of the consistency of nonparametric estimators cannot exist. Additionally, nonparametric identification reveals what types of economic model parameters can be learned from a given type of data, and thus provides a foundation for applied work.

The key strength of our approach is that we develop an identification condition expressed directly in terms of the choice behavior of agents within the model. 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 an agent is characterized by a vector of functions $\theta \in \Theta$, and the economic environment is summarized by $x \in \mathcal{X}$, then the model is given by the relation $y = f(x; \theta)$, where $y \in \mathcal{Y}$ is the agent θ's choice behavior at $x \in \mathcal{X}$.[1]

While the econometrician has data to identify the joint distribution of (y, x) in the underlying population of agents, the agent's characteristics θ are unobservable and heterogeneous among agents. Knowledge of the population distribution G of the unobservable characteristics θ is essential for answering particular economic questions, and it is this distribution that constitutes the target of identification.

[1] Any observable characteristics of an agent are included in the economic environment x.

Choice models of the above form play a prominent role in the applied literature in labor and industrial organization. We develop an identification condition for abstract choice models and show that the condition holds under general conditions in several applied settings. The condition is relatively easy to verify because it is expressed in terms of the decisions that heterogeneous agents make at various choice situations. We focus on applications that extend the relevant model-specific literature in important directions.

We first study the identification of nonadditive random functions (a subject first begun by Matzkin for the case of scalar heterogeneity), where the economic environment $\mathcal{X} \subset \mathbb{R}^K$ is finite dimensional, a type θ is an unknown function $g : \mathcal{X} \to \mathbb{R}^m$, and for any $x \in \mathcal{X}$, $f(x;\theta) = g(x)$. The function g is random from an econometric perspective as it is heterogeneous in the underlying population of agents and cannot be conditioned on by the econometrician. We show that with minimal restrictions on the functional space Θ, we can identify the distribution of marginal effects $Dg(x)$ at any point $x \in \mathcal{X}$. The distribution of marginal effects cannot be observed directly from the data (because the econometrician cannot condition on an individual's g and hence cannot directly observe how any individual responds to changes in x). We also study the full, structural identification of the distribution of g. That is, we identify a distribution over random functions themselves. Thus for example, if we are studying productivity, we identify the distribution over a space of production functions, without limiting attention to a parametric family. For example, we identify the fraction of firms with Cobb-Douglas production functions, the fraction with translog production functions, and so on.

The above results are obtained under the assumption of independence between the regressors x and the structural error term θ. We show that endogeneity can be addressed with instruments and a triangular structure that significantly extends known results on models with endogenous regressors. Extensions of 2SLS to handle heterogeneous, including non-monotone, responses to the instrument are a special case of our results. In a model with a continuous treatment and a continuous instrument for treatment, those who respond more to the treatment can respond more to the instrument for treatment, as might be expected if agents choose treatment levels to maximize their utilities. Further, responses to both the treatment and instrument can be non-monotone: some may increase the treatment intensity as the instrument increases, and some may decrease the treatment intensity.

In industrial organization and marketing, the multinomial choice model is a key tool for demand estimation. In this model, each consumer chooses between J choices. There is a choice specific "special" regressor w_j for each $j \in J$, and a vector v of remaining choice and consumer characteristics. A consumer type θ corresponds to a vector of utility functions $u(v) = (u_1(v), \ldots, u_J(v))$ that give the utility values across choices for all possible values of v. As types are heterogeneous,

the J functions are random across the population. The model is completed by the choice function

$$f\left(\left(v,w\right),\theta\right)=\arg\max_{j\in J}u_{j}\left(v\right)+w_{j}.$$

The outcome in this case is $y=j$, a discrete choice. For full identification of the model, we wish to identify the joint distribution of the J utility functions. Full identification allows the researcher to compute any counterfactual or welfare measure. We allow some elements of v to be endogenous regressors. We discuss extensions to purchases of bundles of items and to the pure characteristics demand model, which weakens support conditions on the w_j's.

All of these cases are applications of our general mathematical framework for identifying distributions of unobserved heterogeneity and can be extended to other economic models (that is, other combinations of what a type θ represents and the form of the choice model $f(x;\theta)$). In a separate paper, we use the framework to explore the identification of the distribution of heterogeneity in selection and treatment effect models, where the selection decision is a multinomial choice (Fox and Gandhi, 2009). In that paper, we present results that extend and generalize those in the literature on selection and treatment effects.

2 The Identification Problem

We consider a general class of economic models where each model $\mathcal{M}$ can be described by a tuple $\mathcal{M}=(\Theta,\mathcal{X},\mathcal{Y},f)$. The set Θ denotes a functional space representing the feasible set of types of agents admitted by the model. The set $\mathcal{X}$ denotes the set of economic environments in the support of the data generating process. The set $\mathcal{Y}$ is the (measurable) outcome space. The function $f:\mathcal{X}\times\Theta\to\mathcal{Y}$ maps an agent's type $\theta\in\Theta$ and economic environment $x\in\mathcal{X}$ to an outcome $y=f(x,\theta)\in\mathcal{Y}$. The joint distribution of outcomes and environments (y,x) is identified from the i.i.d. data. What remains to be identified is the distribution of types $G\in\mathcal{G}$ in the population, where $\mathcal{G}$ is a set of probability measures over Θ.[2]

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

$$\Pr_{G^0}\left(A\mid x\right)=G^{0}\left(\left\{\theta\in\Theta\mid f(x,\theta)\in A\right\}\right)=\int1\left[f(x,\theta)\in A\right]dG^{0}\left(\theta\right). \tag{1}$$

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

[2] In parametric models, the type space Θ is a finite dimensional space. One of the main innovations of the paper is to treat the type space Θ as an infinite dimensional functional space. For the development of the general theory, however, we make no explicit use of any structure on Θ and thus treat it as an arbitrary type space.

the class of such sets $I_{A,x}$ is rich enough to point identify G^0 within a class of distributions $\mathcal{G}$.

To state this problem precisely, let $\Pr(\cdot \mid x)$ be a probability measure over $\mathcal{Y}$ for a given value $x \in \mathcal{X}$ of the environment. Let $P = \{\Pr(\cdot \mid x) \mid x \in \mathcal{X}\}$ denote a collection of such probability measures over all possible economic environments and let $\mathcal{P}$ denote the set of all such collections P. Then we can view (1) as a mapping $L : \mathcal{G} \to \mathcal{P}$. We will say the model $\mathcal{M}$ is *identified* relative to $\mathcal{G}$ if L is one-to-one. That is, $\mathcal{M}$ is identified relative to $\mathcal{G}$ if and only if for any $G, G' \in \mathcal{G}$ and $G \neq G'$, there exists an experiment in the data (A, x) where $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$ such that $\Pr_G(A \mid x) \neq \Pr_{G'}(A \mid x)$, where $\Pr_G(\cdot \mid x)$ and $\Pr_{G'}(\cdot \mid x)$ are the images of G and G' respectively under L.[3]

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

We focus on nonparametric identification, which in our context means that we do not put any parametric structures on either the type space Θ or the set of distributions $\mathcal{G}$.[4] Our main restriction is that we take $\mathcal{G}$ to be the class of all discrete distributions over Θ. Thus the restriction being placed on the distribution of types $G \in \mathcal{G}$ is that the set of types having positive support in the population is at most countable. However the location of the support points and their masses are a priori unknown and need to be identified from the data. Thus $\mathcal{G}$ constitutes an infinite dimensional space of distributions. The class $\mathcal{G}$ can be defined without requiring any a priori structure on Θ, thus allowing us to be fully nonparametric about the type space Θ.[5] As will be demonstrated later, the ability to be fully nonparametric about Θ allows for the general applicability of our results to specific economic contexts.[6]

The focus of this paper is on identification and not estimation. Demonstrating that a model is

[3] In an appendix, 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 straightforward to verify for the models we study in this paper once their identification under the definition presented in the main text has been established.

[4] A lack of nonparametric identification calls into question any parametric estimator of the model: apparently the parametric estimator is only consistent because of parametric functional form restrictions either on the types θ or the distribution G.

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

[6] One defense of the restriction to countable distributions is that the true world is finite. In many of our results, we need to use only variation in $\mathcal{X}$ over the rationals, which is a countable set. In particular, the countable application in Appendix B requires only variation in $\mathcal{X}$ in the rationals. Therefore, we can assume symmetric amounts of variation in θ and x in the true data generating process.

identified does not rule out the ill-posed inverse problem: the inverse of the operator L may not be continuous (in some topology) in the data function P.[7] Thus, nonparametric estimators must adopt some solution to ill-posedness. Bajari, Fox, Kim and Ryan (2009a), or BFKR, present a nonparametric, computationally simple, sieve, linear least squares mixtures estimator for economic choice models. They prove that their estimator of the true distribution G is consistent in the Lévy-Prokhorov metric, if the model is identified. BFKR primarily focus on models where the heterogeneity arises from a finite vector of random parameters rather than an entire function. However, a common function known up to a heterogeneous, finite vector of parameters is a special case of our framework that studies heterogeneous functions. Further, the approach in BFKR could be extended to estimate distributions over functions, with possible additional complexity.[8]

Section 3 presents our identification results for generic economic choice models. Section 4 uses the framework to identify the full distribution of functions in the nonadditve random functions model. Section 5 applies the framework to identify the distribution of marginal effects at a point in the nonadditive random functions model. Section 6 considers multinomial choice models, including those with complementarities across multiple products. Finally, Section 7 investigates nonadditive random functions and multinomial choice models with endogenous regressors. In each of the model-specific sections, we discuss how our results extend the literature on identification for that specific class of model. Appendix A discusses identification with positive probability (in terms of the x's).

3 The Main Result

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

[7] The space of countable 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). Even if a G with countable support is a good distribution to some distribution G' in the space of all distributions, the image of G may be a poor approximation to G' if L is not continuous in G.

[8] Some alternative estimators include 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). Train (2008) considers a series of estimators that rely on the EM algorithm. A large literature in both frequentest and Bayesian statistics considers the estimation of finite and continuous mixtures models with and without covariates (Barbe, 1998; Day, 1969; Roueff and Rydén, 2005). Rossi, Allenby and McCulloch (2005) and Burda, Harding and Hausman (2008) provide flexible Bayesian mixtures estimators for the distribution of random coefficients in the logit and logit-probit models. Hoderlein, Klemelä and Mammen (2008) study the linear regression model with random coefficients. Another 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).

we first define I-sets, which are objects that play a critical role in the remainder of the paper.

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

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

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

Definition 3.2. The model $\mathcal{M}$ is **countably separable** if, for any countable set of types $T \subset \Theta$, there exists a singleton I-set $I^T_{A,x}$.[9]

We now state and prove our main result.

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

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

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

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

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

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

[9] In the definition, $T \subset \Theta$ can be any arbitrary countable subset. The full set of feasible types Θ within the model is typically an uncountably infinite set that is quite distinct from the countable subset T considered in the definition.

implies that

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

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

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

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

The above theorem is properly viewed as an existence theorem, and asserts that under separability of the model, an identifying experiment (A, x) must always exist.[10]

We discuss the identification of distribution of nonadditive random functions in the space of countable distributions in Appendix B. For the main body of the text, we restrict attention to distributions that take on finite, not countable, support (which we refer to as the class of finite distributions). The proof of Theorem 3.3 can be adapted without change for the case where separability applies to finite sets.

Definition 3.4. The model $\mathcal{M}$ is **finitely separable** if, for any finite set of types $T \subset \Theta$, there exists a singleton I-set $I_{Y,x}^T$.

Theorem 3.5. *If the model $\mathcal{M}$ is finitely separable, then the model is identifiable with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

As the proof is identical, we omit it. We learn the number of support points, the identity of support points, and the mass of each support point in identification. As the number of support points of an element of $\tilde{\mathcal{G}}$ can be arbitrarily large, it is not possible to reject the finite support assumption with a finite dataset.[11]

While we have defended the class of distributions $\mathcal{G}$ on the grounds of its sufficient generality, the ideas behind separability can also be applied if we impose the alternative restriction that every

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

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

$G \in \mathcal{G}$ admits a density function. This is discussed in Appendix C. It is important to observe that the class of distributions that admit a density function is not more general than the classes of countable or finite distributions. We provide the argument in Appendix C only to show robustness of the intuition behind separability.

While separability is sufficient for identification, we have not claimed that it is necessary. Teicher (1963) and Yakowitz and Sprangins (1968) investigate the identification of finite mixtures in statistical models without covariates.[12] They show that a necessary and sufficient condition for identification with respect to finite mixtures is that a statistical model satisfies a linear independence property.

The key advantage of separability over the linear independence characterization of identification is that it is more immediately useful. In the context of an economic choice model, linear independence is a non-primitive assumption on the model, and showing linear independence of $\mathcal{M}$ would be equivalent to showing identification itself. The key contribution of separability is that it is expressed in terms of the primitives of an economic choice model and thus can be verified on the basis of the underlying behavior of the agents in the model $\mathcal{M}$ and the variation in the data. We demonstrate the applicability of separability in the remainder of the paper.

3.1 The No Ties Property on Function Spaces

Verifying that a choice model $\mathcal{M}$ satisfies separability is closely related to the underlying functional space Θ satisfying a "no ties" property that we formalize below. There are two versions of the no ties property, both a strong and a weak version, and both properties are satisfied by functional spaces that are quite commonly used in economic models. We use one version or the other of the assumption in all sections of the paper except the section on identifying marginal effects, which operates in a more general function space. To establish some notation, for a given non-empty rectangle $\mathcal{X} \subseteq \mathbb{R}^k$, let $\mathcal{C}_{\mathcal{X}}^{k,m}$ denote the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}^m$.

Definition 3.6. A set of functions $\mathcal{F}_{\mathcal{X}}^{k,m} \subseteq \mathcal{C}_{\mathcal{X}}^{k,m}$ satisfies the strong no ties property (SNTP) if for any finite subset of functions $\{g_1, \ldots, g_n\} \subset \mathcal{F}_{\mathcal{X}}^{k,m}$ and any open $U \subseteq \mathcal{X}$, there exists a point $x \in U$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$.

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

[12] Blum 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.

the set $\mathcal{A}_{\mathcal{X}}^{k,m}$ of vector valued real analytic functions (which contains $\mathcal{P}_{\mathcal{X}}^{k,m}$) satisfies the the SNTP. See Appendix D for a proof.[13] Even more generally, we can apply Zorn's lemma to produce a maximal set of functions $\mathcal{S}_{\mathcal{X}}^{k,m} \subset \mathcal{C}_{\mathcal{X}}^{k,m}$ that satisfies the SNTP and contains $\mathcal{P}_{\mathcal{X}}^{k,m}$ as a subset. In the applications to follow, we will use this maximal set $\mathcal{S}_{\mathcal{X}}^{k,m}$ as the functional space satisfying the SNTP.

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

Definition 3.7. A subset $\mathcal{F}_{\mathcal{X}}^{k,m} \subseteq \mathcal{C}_{\mathcal{X}}^{k,m}$ satisfies the weak no ties property (WNTP) if for any finite subset $\{g_1, \ldots, g_n\} \subset \mathcal{F}_{\mathcal{X}}^{k,m}$ there exists $x \in \mathcal{X}$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$.

Once again Zorn's lemma applies and there exists a maximal subset $\mathcal{W}_{\mathcal{X}}^{k,m} \subset \mathcal{C}_{\mathcal{X}}^{k,m}$ that satisfies the WNTP and contains the class of polynomials as a subset.[14] In the applications to follow, we will use the maximal set $\mathcal{W}_{\mathcal{X}}^{k,m}$ as the function space satisfying WNTP. Observe that by construction $\mathcal{S}_{\mathcal{X}}^{k,m} \subseteq \mathcal{W}_{\mathcal{X}}^{k,m}$.

In the remainder of the paper, we show finite separability and hence identification of choice models $\mathcal{M}$ with respect to the finite distributions $\tilde{\mathcal{G}}$ by exploiting the SNTP or the WNTP on the underlying functional space of types Θ. Showing identification with respect to countable distributions would analogously proceed by establishing countable versions of the WNTP and SNTP. We show an example of just such a result for the case of identifying countable distributions over the space of non-additive random function in Appendix B.

[13] Real analytic functions are defined formally in the appendix, but roughly speaking, they are functions that can be parameterized by a countable parameter vector. The previous literature studies identification of the distribution of random coefficients in the linear regression model. The linear regression model nests polynomials of an a priori fixed order. The space of real analytic functions nests all polynomials of any finite order as well as polynomials of countable order. Examples of real analytic functions include the simple functions such as exp, sin, and log, as well as algebraic combinations and compositions of these functions. Commonly used production and demand functions, such as the translog, are real analytic.

[14] Zorn's lemma states that for any partially ordered set, if every chain has an upper bound, then the set has at least one maximal element. To see the applicability of Zorn's to existence of a maximal set of functions satisfying the SNTP or the WNTP, let us consider the WNTP (the SNTP argument follows similarly). Let $P\left(\mathcal{C}_{\mathcal{X}}^{k,m}\right)$ denote the power set (the set of all subsets) of continuous functions from the non-empty rectangle $\mathcal{X} \subseteq \mathbb{R}^k$ to $\mathbb{R}^m$, and consider the set of sets $W = \left\{A \in P\left(\mathcal{C}_{\mathcal{X}}^{k,m}\right) \mid A \text{ satisfies the WNTP}\right\}$. W is partially ordered under the subset relation $\subseteq$, and consider any chain $D \subset W$ (a totally ordered subset). The set $\cup_{A \in D} A$ is an upper bound for D under the order $\subseteq$, and hence we need only show that $\cup_{A \in D} A \in W$, i.e., that it satisfies the WNTP. Consider any finite set of functions $\{g_1, \ldots, g_n\} \in \cup_{A \in D} A$. By the fact that a chain is totally ordered, there exists a $A^* \in D$ such that $\{g_1, \ldots, g_n\} \subset A^*$. Because A^* satisfies the WNTP, we can find $x \in \mathcal{X}$ such that $g_i \neq g_j$ implies $g_i(x) \neq g_j(x)$. Hence $\cup_{A \in D} A \in W$.

4 Identifying Distributions over Nonadditive Random Functions

The most basic choice model we consider, which generalizes the nonparametric regression model, is identification over nonadditive random functions. In this model, the economic environment is summarized by $x \in \mathcal{X} \subset \mathbb{R}^m$. A type θ is a function $g : \mathcal{X} \to \mathbb{R}^m$, and the choice model is such that g's choice behavior at $x \in \mathcal{X}$ is $f(x, g) = g(x)$. The technical assumptions on the model are as follows.

Assumption 4.1. *$\mathcal{X}$ is a non-empty rectangle.*

In this paper, we allow only continuous covariates in $\mathcal{X}$. If discrete characteristics d exist, we can condition on them. In other words, we can identify a distribution $G(g \mid d)$ over functions $g(x \mid d)$ for each observable value of d.

Assumption 4.2. *Economic environments $x \in \mathcal{X}$ are distributed independently of types $g \in \Theta$.*

Assumption 4.3. *The type space Θ is equal to the $\mathcal{W}_{\mathcal{X}}^{k,m}$.*[15]

In the standard nonparametric regression model that is typically taken to economic data, if an agent's choice variable is a multivariate outcome $y \in \mathbb{R}^m$, then heterogeneity among agents can be summarized by a finite dimensional vector $\epsilon \in \mathbb{R}^m$, and furthermore it is often assumed that the choice model is $y = f(x, \epsilon) = f(x) + \epsilon$. The structural error term ϵ in the nonparametric regression model is thus both finite dimensional and enters the choice model in an additively separable fashion. In the present model, the structural error term g is neither finite dimensional (indeed it has support in an infinite dimensional functional space $\mathcal{W}_{\mathcal{X}}^{k,m}$) nor does it enter the choice model f in an additive way. A key example is identifying the distribution of production functions among firms in an industry. Abstracting for the moment from the problem of endogeneity in a firm's choice of inputs (an issue we address later in the paper), if we observe variation in the input choices across firms, the present section shows that we can nonparametrically identify a distribution over an infinite-dimensional space of production functions. For example, we can recover the fraction of firms with Cobb-Douglas production functions, translog production functions, and so forth. In order to identify production functions through the lens of the traditional nonparametric regression model, it must be assumed that all firms have the same underlying technology (as captured through

[15] We explore identification using a local space $\mathcal{X}$. The ability to identify a distribution of functions using local support arises from the WNTP. In principal, we could split the master space $\mathbb{R}^k$ into many disjoint subsets $\mathcal{X}$. One could apply the identification arguments in this paper for each subset $\mathcal{X}$ separately. The maximal set $\mathcal{W}_{\mathcal{X}}^{k,m}$ will vary with $\mathcal{X}$.

$f(x)$), and all heterogeneity amongst firms must occur along a Hicks neutral productivity dimension (as captured through the scalar total factor productivity ϵ).

Using our main result Theorem 3.5, the proof of identification in the present case is straightforward. For notational ease, we introduce the following shorthand notation that will be used in the remainder of the paper to define a particular form of an I-set. For any $z \in \mathbb{R}^m$, let $A_z = \{y \in \mathbb{R}^m \mid y \leq z\}$. We will use the simpler I-set notation $I^T_{z,x}$ to denote the I-set $I^T_{A_z,x}$. Thus for any subset of types $T \subset \Theta$, $I^T_{z,x} = \{g \in T \mid g(x) \leq z\}$.

Theorem 4.4. *Under Assumptions 4.1, 4.2, and 4.3, the distribution of nonadditive random functions $G(g)$ is identified with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

Proof. We show that the model satisfies finite separability. Thus take any finite subset of types $T = \{g_1, \ldots, g_n\} \subset \Theta$. We now show how to produce a singleton I-set of the form $I^T_{z,x}$. As $\Theta = \mathcal{W}^{k,m}_{\mathcal{X}}$ is assumed to satisfy the WNTP, then there exists an $x \in \mathcal{X}$ for which all functions in T take distinct values. For this $x \in \mathcal{X}$, let z denote a minimal element from the set $\{g_i(x) \mid i = 1, \ldots, n\}$, where the order is the standard partial order on $\mathbb{R}^m$ given by $\leq$. A minimal element always exists by the finiteness of the set. By the choice of x, there is a unique $j \in \{1, \ldots, n\}$ for which $g_j(x) = z$, and because z is minimal, $I^T_{z,x} = \{g_j\}$, and thus we have a singleton. □

4.1 Literature Review for Nonadditive Random Functions

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 lies in an infinite dimensional space that includes the space of polynomials and real analytic functions. We know of no other work that attempts to identify a nonparametric distribution over an infinite dimensional, nonparametric class of functions in the context of nonparametric regression. We discuss the extension of Theorem 4.4 to endogenous regressors in a later section.

Matzkin (2003) studies the identification of models of the form $y = f(x, \theta)$ where f is an unknown function that is common across agents and θ is an unobservable scalar that varies across agents. Matzkin considers three identification conditions, including when $f(x, \theta)$ is restricted to be monotone in θ. We study models of the form $y = f(x, \theta) = g(x)$ where $\theta = g$ is an unknown function that varies across agents. Our notation drops the distinction between the homogeneous function f and the scalar heterogeneous disturbance θ in Matzkin's notation. In our notation, each agent has its own function. We identify a distribution over an infinite-dimensional space rather than one infinite-dimensional function and a distribution over a scalar.

5 Identifying Distributions of Marginal Effects

In the previous section, we nonparametrically identified a distribution over random functions $g(x)$ from data $F(y \mid x)$ on the conditional distributions of choices y given environments x. Having identified the distribution G over random functions $g \in \mathcal{W}_{\mathcal{X}}^{k,m}$, any counterfactual of interest can be computed. For example, the distribution of the treatment effect $g(x_1) - g(x_0)$ can be derived from knowledge of G. Knowledge of $F(y \mid x)$ by itself without identification of G is only sufficient to identify the average treatment effect (ATE) $E[y \mid x_1] - E[y \mid x_0] = E_g\left[g\left(x_1\right) - g\left(x_0\right)\right]$, because of the linearity of expectations. Unless the treatment effect is homogeneous for all members of the population, which is implied by the standard additive representation of heterogeneity in the nonparametric regression model $y = f(x)+\epsilon$, then more policy information is learned by identifying G from $F(y \mid x)$.

In some cases, the policy counterfactual of interest is the treatment effect associated with a marginal change in x, namely the distribution of marginal effects $Dg\left(x^*\right)$ for some specified $x^\star$, where $Dg\left(x\right)$ is the derivative of the function $g : \mathcal{X} \to \mathbb{R}^m$ at an interior point $x \in \mathcal{X}$. Recall that the derivative of a multivariate function from $\mathbb{R}^k$ to $\mathbb{R}^m$ at a point x is a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^m$ that can be represented by the Jacobian matrix

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

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

If the distribution of the marginal treatment effect is the policy counterfactual of interest, then rather than seek identification over random functions g, which is sufficient for identifying the policy counterfactual, we can seek identification of the distribution of marginal effects directly. This more limited identification question allows us to relax altogether the assumption made in the previous section that the type space Θ satisfies the WNTP. Thus suppose that the counterfactual of interest is the distribution over the marginal treatment effect J_{g,x^*} at an interior point $x^* \in \mathcal{X}$. Let the underlying type space Θ denote all functions from $\mathcal{X}$ to $\mathbb{R}^m$ that are differentiable at x^*. Observe that within Θ, there exist types $g \neq g'$ that differ from each other globally (there exist a $z \in \mathcal{X}$ such that $g(z) \neq g'(z)$) but have the same local behavior at x^* ($g(x^*) = g'(x^*)$ and $Dg(x^*) = Dg'(x^*)$).

From a policy perspective that is concerned with the distribution of marginal effects at x^*, the distinction between g and g' is not policy relevant.

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

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

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

Lemma 5.1. *For any finite set of functions $g_i : \mathcal{X} \to \mathbb{R}^m$ for $i = 1, \ldots, n$ that are differentiable at $x^* \in \mathcal{X}$, if $g_i(x^\star) = g_j(x^\star)$ and $Dg_i(x^\star) \neq Dg_j(x^\star)$ for all $i \neq j$, then for any ball $B_\epsilon(x^\star)$ with $\epsilon > 0$, there exists a $x^\epsilon \in B_\epsilon(x^\star)$ such that $g_i(x^\epsilon) \neq g_j(x^\epsilon)$ for all $i \neq j$.*

The proof is Appendix E. The lemma does not require the WNTP or the SNTP, which is why the following theorem uses assumptions weaker than Theorem 4.4.

Theorem 5.2. *Under Assumptions 4.1 and 4.2, the distribution H over the type space $\Theta_{x^\star}$ is identified in the class of finite distributions $\tilde{H}$. That is, the distribution of marginal effects at $x^\star$ is identified.*

Proof. The proof verifies verifies finite separability of the model. Consider a finite subset of types $T = \{g_1, \ldots, g_N\} \subset \Theta_{x^*}$. There are two cases to consider.

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

The second case is when multiple types take on the minimal value $y^\star$ at $x^\star$, and thus $I^T_{y^*,x^*}$ is not a singleton set. Observe that since T is finite and since each $g \in T$ is continuous, there exists an $\epsilon > 0$, say $\bar{\epsilon}$, such that $x \in B_{\bar{\epsilon}}(x^*)$ implies that for $g \in I^T_{y^*,x}$ and $g' \in T - I^T_{y^*,x^*}$, $g'(x) \not\leq g(x)$ (since by construction $g'(x^*) \not\leq g(x^*)$). In addition, observe that for any pair of functions g_i and g_j in $I^T_{y^*,x^*}$, $g_i(x^\star) = g_j(x^\star)$ but $Dg_i(x^\star) \neq Dg_j(x^\star)$. Thus by Lemma 5.1, for any $\epsilon > 0$, there exists a $x^\epsilon \in B_\epsilon(x^\star)$ such that $g_i(x^\epsilon) \neq g_j(x^\epsilon)$ for all pairs of functions g_i and g_j in $I^T_{y^*,x^*}$. Choose

$\epsilon > 0$ small enough so that for any $x \in B_\epsilon(x^\star)$, $x \in \mathcal{X}$ and $\epsilon < \bar{\epsilon}$. Then for any $x \in B_\epsilon(x^\star)$, there exists a minimal element y^* of the set $\{g_1(x), \ldots, g_N(x)\}$ that is attained by a unique type, $y^* = g_i(x)$ for a unique type $g_i \in T$.[16] Thus $I^T_{y^*,x}$ is a singleton consisting of only g_i. □

5.1 Literature Review for Marginal Effects

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

6 Multinomial Choice

Multinomial choice is a key model used in empirical industrial organization to model consumer demand. Demand functions are useful for measuring market power and predicting the welfare gain from new goods. This section shows how discrete choice models of demand are nonparametrically identified within our framework. Furthermore, we differentiate our framework from the existing literature on identification in multinomial choice by showing that we are able to relax the large support assumptions on the "special regressor" that has now become standard in the literature.

6.1 Base Case for Multinomial Choice

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

Assumption 6.1. *Let $V \subset \mathbb{R}^K$, the support of v, be a non-empty rectangle. Let $x = (v, w) \in \mathcal{X} = V \times W_1 \times \cdots \times W_J$ where $W_j = \mathbb{R}$ for each $j \in J$.*

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

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

Assumption 6.2. *The function u is statistically independent of the observable choice set $x = (v, w)$.*

Furthermore, to show separability, we will need a monotonicity assumption for the special regressor w_j.

Assumption 6.3. *The utility of a type u purchasing product j is $u^j(v) + w_j$.*

The additive separability of w_j ensures that at any v there will be a set of w_j's where a given type u will switch to a different choice. We also introduce an outside good that we label good $j = 0$ whose utility is normalized to 0 for each agent. An agent's response at $x = (v, w)$ is given by

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

where $u^0(v) + w_0 \equiv 0$. Further, we enforce the partial tie-breaking rule that if $\arg\max_{j \in \mathcal{Y}} \left\{u^j(v) + w_j\right\} = 0$, then the outside good is chosen.[17] We restrict attention to utility functions that satisfy the weak no-ties property.

Assumption 6.4. *The type space Θ of feasible utility functions is equal to the weak no tie breaking set $\mathcal{W}_V^{K,J}$.*

A few comments on the model are in order. A special case of this framework is when only the components of v corresponding to product j enter $u^j(v)$: $u^j(v) = u^j(v_j)$.[18] Letting the utility to product j also depend on the characteristics of products $k \neq j$ can capture the idea of context or "menu" effects in consumer choice. Even if such effects are not economically desirable, there is no cost to us in mathematical generality and thus we let the whole menu enter as an argument to each u^j. The choice-specific scalar w_j, however, enters preferences in an additively separable way (and hence preferences are quasilinear in this scalar characteristic). One example is that w_j could be the price of good j, in which case $u^j(v)$ is type u's reservation price for product j, and preferences are better expressed as $u^j(v) - w_j$. However, w_j could be some non-price product characteristic or, with individual data, an interaction of a consumer and product characteristic, like the geographic distance between a consumer and a store.

[17] The tie breaking rule is not essential to the overall identification argument. Once an identifying experiment has been found using this tie breaking rule, we can find another identifying experiment that does not depend upon any particular form of the tie breaking rule.

[18] An even more typical empirical specification is when an agent u's sub-utility functions u^j are the same across $j = 1, \ldots J$, and each agent also receives a product-specific "idiosyncratic" error term ϵ^j. This is a special case of the framework we consider.

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

We will later discuss the "large support" assumption on the support of each w_j and how the assumption's role contrasts with the role it plays in other approaches to identification in multinomial choice. For now we have the main result.

Theorem 6.5. *Under assumptions 6.1, 6.2, 6.3, and 6.4, the distribution of utility functions in the multinomial choice model is identified with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

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

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

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

According to Definition 3.7, there exists a $v \in V$ such that $u_i(v) \neq u_j(v)$ for all $u_i \neq u_j$, $u_i, u_j \in T$. Because the vector of $u(v)$'s at $v \in V$ for $u \in T$ is finite, there exists a minimal vector $u_i(v)$. By minimal vector, we mean $u^j_k(v) > u^j_i(v)$ for some $j \in \mathcal{Y} - \{0\}$, $\forall\, u_k \neq u_i$, at v. There could be multiple minimal vectors; we focus on one. Then set the vector $-w = u_i(v)$. This means that the vector of product specific utilities $u_i(v) + w = 0$ for type u_i. By the tie breaking rule, type $u_i(x)$ purchases the outside good. All other types $u_j \in T - \{u_i\}$ purchase an inside good at $x = (v, w)$, as $u^j_k(v) > u^j_i(v)$ for some $j \in \mathcal{Y} - \{0\}$, $\forall\, u_k \neq u_i$, at v. Thus, $I^T_{0,v,w} = \{u_i\}$. □

A major difference relative to the simultaneously-developed literature is that we identify the full joint distribution of J utility *functions* $\left\{u^j(\cdot)\right\}_{j=1}^J$. For example, Berry and Haile (2008) identify a distribution $F_t(\cdot \mid v)$ of utility *values* $t = (t_1, \ldots, t_J)$ conditional on v, where $t_j = u_j(v)$ for a particular v. Identifying an unconditional distribution of utility functions rather than a conditional distribution of utility values has several uses in structural empirical work. For example, utility functions can be used to compute the utility differences of particular structural types u at old and new choice sets. For example, our theorem allows us to compute the joint distribution of $\left\{u^j(v') + w'_j - u^j(v) - w_j\right\}_{j=1}^J$, the utility improvement for each of the J products if choice sets

or observable consumer characteristics in $x = (v, w)$ change. We can also calculate the distribution of

$$\left\{ \arg\max_{j \in \mathcal{Y}} \left\{ u^j (v') + w'_j \right\} - \arg\max_{j \in \mathcal{Y}} \left\{ u^j (v) + w_j \right\} \right\}, \tag{4}$$

the differences in maximized utility values, one version of a "treatment effect" for changing (v, w) to (v', w'). By contrast, the distribution $F_t (t_1, \ldots, t_J \mid v)$ does not assign utility to particular structural types, and so a researcher cannot calculate (4). The lack of utility functions prevents the researcher from computing a distribution of welfare changes, a major use of structural demand models.[19]

6.2 Support Conditions on the Special Regressors w

An alternative identification strategy in multinomial choice is to vary the vector of special regressors $(w_1, \ldots, w_J)$ so as to "trace" the CDF of the underlying distribution of latent utility values $(t_1, \ldots, t_J) = \left(u^1(v), \ldots, u^J(v)\right)$ conditional on v (Matzkin, 1993; Lewbel, 2000; Berry and Haile, 2008). More precisely, fixing the product characteristics $v \in V$, this literature considers identification of the joint distribution of the latent utilities $(t_1, \ldots, t_J)$ by tracing the CDF through the relationship

$$\Pr (j = 0 \mid w_1, \ldots, w_J, v) = F_t (-w_1, \ldots, -w_J \mid v),$$

where the random vector $t = (t_1, \ldots, t_J)$ has a joint distribution characterized by the conditional CDF $F_t(\cdot \mid v)$. Thus using variation of the special regressors $w = (w_1, \ldots, w_J)$ over all of $\mathbb{R}^J$ while holding fixed the value of $v \in V$ enables identification of the conditional CDF $F_t (\cdot \mid v)$ for all $v \in V$.

A problem with the tracing-the-CDF approach is the requirement that the scalar characteristics $(w_1, \ldots, w_J)$ have full support over $\mathbb{R}^J$, and hence these characteristic have acquired the title of "special regressors" in the literature. Of course, if the researcher restricts utilities $u_j (v)$ so that they are bounded a priori for any v, say between $[-m, m]$, then w_j would only require variation between $[-m, m]$. Thus "large support" more specifically refers to the requirement that the support of $(w_1, \ldots, w_J)$ covers the support of the latent utilities $(t_1, \ldots, t_J)$ for any value of the product characteristics $v \in V$. Unfortunately, there does not exist any natural way to bound the support of $(t_1, \ldots, t_J)$ for even a fixed $v \in V$. Hence the support requirement on $(w_1, \ldots, w_J)$ cannot be shrunk beyond $\mathbb{R}^J$ if the CDF is to be traced using $(w_1, \ldots, w_J)$ for each $v \in V$.

We now show (for the first time to our knowledge) that a nonparametric random utility model can be identified with an arbitrarily small support on the special regressors by exploiting restrictions from economic theory (that is, adding a restriction on preferences to the above general model of demand). We achieve this result using separability and, as we show, the same result could not be

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

attained by an approach to identification that tries to trace the CDF, thus distinguishing the role that the special regressor plays in the two contexts. The application also highlights an important advantage of identification via separability: as separability is a primitive of the economic model itself, it can more easily incorporate theoretical restrictions on the model to aid with identification.

The particular economic restriction that we draw upon is a variant of the "pure characteristics" demand model (Bajari and Benkard, 2005; Berry and Pakes, 2007), which assumes that all types $u \in \Theta$ value products for their characteristics v and not because of an idiosyncratic taste shock ϵ^j.

Assumption 6.6. *Let $V \subset \mathbb{R}^K$, the support of v, be a non-empty rectangle. Let $x = (v, w) \in V \times W_1 \times \cdots \times W_J$ for $W_j = [-\delta_j, \delta_j]$ for some scalar $\delta_j > 0$ for each $j \in J$.*

Thus the special regressors have arbitrarily small support. We now strengthen the structure on the space of utility functions.

Assumption 6.7. *The type space Θ of utility functions is a subset of of the no tie breaking set $\mathcal{S}_{\mathcal{V}}^{K,J}$ that satisfies the following: there exists an interior point $v^0 \in V \subseteq \mathbb{R}^K$ such that for all $u \in \Theta$, $u^j\left(v^0\right) = 0$ for all $j \in J$.*

The product characteristics v^0 correspond to a menu of characteristics in which all of the inside goods are identical to the outside good (and hence all agents value them identically). Thus for any menu of characteristics v inside a small ball $B_\epsilon\left(v^0\right)$ around v^0, all inside goods are "similar" to both each other and the outside good. If the model $\mathcal{M}$ admits such an $v^0 \in V$, then we refer to it as a pure characteristics demand model.

Importantly, identification will not require large support on either v or w.

Theorem 6.8. *Under assumptions 6.2, 6.3, 6.7, and 6.6, the joint distribution of utility functions in the pure characteristics multinomial choice model is identified with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

Proof. We verify finite separability. Let a finite $T \subseteq \Theta$ be given. For each $u \in T$, by continuity there exists ϵ_u such that $v \in B_{\epsilon_u}\left(v^0\right)$ implies $\left|u^j(v)\right| < \delta$ for all $j \in \mathcal{Y}$. Take $\epsilon = \min_{u \in T} \epsilon_u$. Also, by the definition of the SNTP, there exists $v \in B_\epsilon(v^0)$ such that $u_i(v) \neq u_j(v)$ for all $u_i \neq u_j$, $u_i, u_j \in T$. Then the remainder of the proof of Theorem 6.5 can be used to complete the argument. □

Thus under the pure characteristics assumption, so long as there exists product characteristics in the support of the data generating process that are arbitrarily close to the point at which all goods are identical (in characteristics) to the outside good, we can achieve identification via separability. The main purpose of this demonstration is to distinguish the role of the special regressor in the proof of separability form the role of the special regressor in tracing the CDF. The underlying population of latent utilities $\left(u^1(v), \ldots, u^J(v)\right)$ cannot be bounded for all $u \in \Theta$

when $v \neq v^0$, and hence the support requirement needed to trace the CDF cannot be shrunk from $W_j = \mathbb{R}$ for any such v. Using separability, however, the economic restrictions implicit in the pure characteristics model can be used as information that allows us to substantially relax the support requirement on the special regressor.

6.3 Purchasing Multiple Products with Complementarities or Substitutes in Preferences

Gentzkow (2007) and Liu, Chintagunta and Zhu (2008) study choice situations where each discrete choice $j = 0, \ldots, J$ indexes a bundle of composite choices. For example, a consumer can purchase cable television separately ($j = 1$), purchase an internet connection separately ($j = 2$), purchase both cable television and an internet connection together as a bundle ($j = 3$), or purchase nothing, the outside good ($j = 0$). The goal in this situation is to distinguish between explanations for observed joint purchase: are consumers observed to buy cable television and an internet connection at the same time because those who watch lots of television also have a high preference for 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 $G(u)$ that gives positive correlation between the utility functions $u^1(v)$, $u^2(v)$, and $u^3(v)$. True complementarities are measured by

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

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

As we have already explored in Theorems 6.5 and 6.8, we can identify the joint distribution of heterogeneity, which means we can identify the distribution of complementarities as a function of the joint distribution $G(u)$, if prices w_j are bundle-specific. Thus, we need to observe different choice situations where the bundle is or is not aggressively priced relative to the singleton packages. This is the data scheme for Liu et al.: they observe different bundles of telecommunications services at different prices, across geographic markets.

6.4 Literature Review for Multinomial Choice

Matzkin (2007) surveys the literature on heterogeneous choice, emphasizing the scarcity of results on discrete choice models about the nonparametric identification of the distribution of heterogene-

ity, the distribution G of u, even though random coefficients are a critical tool in the empirical literature. Even papers that emphasizes the flexibility of a particular specification for heterogeneity do not formally prove identification (McFadden and Train, 2000; Rossi and Allenby, 2003; Burda et al., 2008).[20]

Briesch, Chintagunta and Matzkin (2009) study the identification of a discrete choice model where the payoff to choice j is $V(j, s, v_j, \omega) + \epsilon_j$, 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 (2009b) were the first to prove the identification of the random coefficients logit model with continuous characteristics. They use calculus to show that all of the moments of the random coefficients are identified. The proof relies on linearity, $u^j = x_j'\beta$, but, unlike other work, only variation in $x_j'\beta$ around the value $x_j = 0$ is needed. Neither of the papers above deal with endogenous regressors.

Some differences with the paper by Berry and Haile (2008) are mentioned above. We discuss this paper below in the section on endogenous regressors, as well. Chiappori and Komunjer (2009) discuss some assumptions under which they can show the identification of a multinomial choice model without additive regressors. Manski (2007) considers the identification of a counterfactual choice function when there is a fixed number of decision problems x and hence a fixed number of types with different responses at those x's. He also imposes independence between choice sets x and preferences and focuses on set identification. We point identify a distribution of utility functions on the space of all functions satisfying the WNTP, Definition 3.7.

Studying the special case of $J = 1$, one inside good and one outside good, Ichimura and Thompson (1998) use the Cramér and Wold (1936) theorem for identification, which relies critically on a linear index functional form: $u^j(v, w) = v_j'\beta + w_j$. We use only the quasilinearity of $u^j(v) + w_j$ in w_j and the WNTP. A space of linear functions distinguished by the parameter β trivially satisfies the WNTP. A key assumption in both papers is monotonicity in at least one regressor w_j. Ichimura and Thompson also need full support on all covariates (both v and w) to apply the Cramér-Wold theorem. Further, Ichimura and Thompson need an identification condition that reduces to our monotonicity condition that the sign of w_j in $u^j(v) + w_j$ is known. We need large support on only w in Theorem 6.5. Gautier and Kitamura (2007) provide some alternative identification arguments (the results are the same) and a computationally-simpler estimator for the model of Ichimura and Thompson.

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

7 Endogenous Regressors

7.1 Endogenous Regressors and Nonadditive Random Functions

Endogenous regressors are often encountered in social-science applications. Thus in the context of nonadditive random functions, where a type is a mapping $g : \mathbb{R}^K \to \mathbb{R}^M$, it is possible that some subset of the regressors, say the first $J < K$ regressors are not stochastically independent of an agent's type $g \in \Theta$ (due perhaps to endogenous sorting into x's). Denote the first J regressors (the endogenous regressors) as $\tilde{x} = (\tilde{x}_1, \ldots \tilde{x}_J)$ and the last $N = K - J$ regressors (the exogenous regressors) as $x = (x_1, \ldots x_N)$.

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

To handle the endogeneity problem, we extend the use of instrumental variables to allow for both heterogeneity in the primary economic equation (that is heterogeneity in random functions g), along with heterogeneity in how a type responds to the instrument. That is, we treat the IV equation as a non-additive random function as well. In particular, we assume that there exists a vector of instruments $z = (z_1, \ldots, z_J) \in Z \subset \mathbb{R}^J$ that are independent of the type g and that along with the exogenous regressors $x \in X$ determine the endogenous regressors through an auxiliary equation $\tilde{x} = h(x, z)$.[21] A type consists of a pair of functions (g, h) , and the choice model in turn can be expressed as a recursive system of equations. For an economic environment $(x, z) \in X \times Z$ and type (g, h), the choice model $f((x, z), (g, h))$ predicts two outcomes, namely $\tilde{x} \in \mathbb{R}^J$ and $y \in \mathbb{R}^M$, where

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

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

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

The essential feature of the model is that the exogenous variables (x, z) are stochastically independent of the type (g, h), although the distribution of g can depend on $\tilde{x}$ conditional on the exogenous regressors (x, z), which is the source of the endogeneity problem. A special case of this model is linear 2SLS where all of the coefficients in both the outcome and IV equations are random with potential joint dependency in the coefficients across equations. That is, the random coefficients in the primary equation have an unrestricted joint distribution with the random coefficients in the IV equation.

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

Assumption 7.1. *Let the support of the exogenous variables (x, z) be the Cartesian product $X \times Z$, where $X \subseteq \mathbb{R}^N$ and $Z \subseteq \mathbb{R}^J$ are both non-empty rectangles.*

We impose the following restriction on the functional space of types, which requires that a type (g, h)'s outcome equation g lie in a functional space satisfying the SNTP and that the IV equation h lie in a functional space satisfying the WNTP, and further that the IV equation is capable of "moving around" the endogenous variables in a sense we make formal below.

Assumption 7.2. *The type space Θ is a subset of $\mathcal{S}^{K,M}_{\mathbb{R}^K} \times \mathcal{W}^{K,J}_{X \times Z}$ such that the following conditions hold: (i) For any type $(g, h) \in \Theta$, the derivative $D_z h(x, z)$ of a type's IV equation with respect to the instruments everywhere exists in the interior of $X \times Z$, and is continuous in $(x, z) \in X \times Z$; and (ii) For any type $(g, h) \in \Theta$, and for any interior $x \in X$, the derivative $D_z h(z, x)$ 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 powerful instrument almost everywhere. For any type $(g, h) \in \Theta$, almost everywhere local variation in z can induce the endogenous regressors $(\tilde{x}_1, \ldots, \tilde{x}_J)$ to vary locally in a full rank way, holding the exogenous regressors x fixed. Thus fixing $x \in X$ and for almost all $z \in Z$, the local variation in $\tilde{x}$ induced by the local variation in z is not restricted to a lower dimensional subspace.

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

Assumption 7.3. *The type (g, h) is stochastically independent of the instruments and exogenous regressors (x, z).*

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

Theorem 7.4. *Under assumptions 7.1, 7.2, and 7.3 the distribution of nonadditive random functions (g, h) with endogenous regressors is identified with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

Proof. We proceed by showing finite separability of the model. Thus we take an arbitrary finite set of types $T \subset \Theta$ and seek to construct a singleton I-set. To fix I-set notation, observe that the choice variables of the model are $(y, \tilde{x}) \in \mathbb{R}^{M+J}$ and the exogenous variables are $(x, z) \in X \times Z$. Hence for any finite set of types $T \subset \Theta$, we consider I-sets that take the form[22]

$$I^T_{(y,\tilde{x}),(x,z)} = \{(g,h) \in \Theta \mid h(x,z) = \tilde{x} \,\text{and}\, g\left(h(x,z),x\right) = y\}\,.$$

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

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

In case 2, we have that the set $T_3 = \left\{g \in \mathcal{S}^{K,M}_{\mathbb{R}^K} \mid (g,h_1) \in T_2\right\}$ is not a singleton. Observe that by Assumption 7.2, , we can find a $z^* \in Z$ such that $(x,z^*) \in U$ and the Jacobian $D_z h_1(x,z^*)$ has full rank J. Furthermore, by continuous differentiability of h_1, the Jacobian $D_z h_1(t_1,t_2)$ has full rank J for all (t_1,t_2) in a sufficiently small ball $V \subseteq U$ containing (x,z^*).

As a consequence of the Jacobian having full rank everywhere in V, the change of variable mapping $(x,z) \mapsto (h(x,z),x)$ defined over V, which we denote by R, is an open mapping by consequence of the open mapping theorem,[23] and thus the image $R(V)$ is an open set in $\mathbb{R}^K$. Now using the SNTP, there exists $(x',z') \in V \subseteq X \times Z$ such that for all distinct functions g_i and g_j in T_3, $g_i\left(h_1\left(x',z'\right),x'\right) \neq g_j\left(h_1\left(x',z'\right),x'\right)$. We can now repeat the argument from case 1 to generate a singleton I-set. That is, we can pick any point y^* from the set of values $\{g(h_1(x',z'),x') \mid g \in T_3\} \subset \mathbb{R}^M$, and observe that by construction y^* is attained at a unique $g \in T_3$, which we can denote as g_1. Then observe the I-set $I^T_{(y^*,h_1(x',z')),(x',z')}$ is a singleton consisting of only the type (g_1,h_1). □

[22] In terms of the main Theorem 3.5, we are considering measurable subsets in the choice outcome space of the form $A_{y,\tilde{x}} \subset \mathbb{R}^{M+J}$, where $A_{y,\tilde{x}}$ is a singleton set $\{(y,\tilde{x})\}$. We are thus using the notation $I^T_{(y,\tilde{x}),(x,z)}$ as shorthand for what is more formally expressed as $I^T_{A_{y,\tilde{x}},(x,z)}$.

[23] The matrix of partial derivatives of R is of the form $A = \begin{bmatrix} D_x h(z,x) & I_N \\ D_z h(z,x) & 0_{J,N} \end{bmatrix}$, 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 $D_z h(z,x)$ is invertible. Therefore, by the open-mapping theorem, $(x,z) \mapsto (h(x,z),x)$ is an open mapping.

7.2 The Generality of the Identification Result for Endogenous Regressors

The generality of the identification argument we have just proved should not be lost in the notation. A very special case of Theorem 7.4 is showing identification for a linear IV model, 2SLS, with random coefficients in both the first stage and the outcome equation. Let y, $\tilde{x}$, x and z all be scalars for exposition. The a type (h, g) is a system of equations

$$\begin{aligned} \tilde{x} &= a_0 + a_x x + a_z z \\ y &= b_0 + b_x x + b_{\tilde{x}} \tilde{x}, \end{aligned} \tag{5}$$

where a type θ can be represented as the unknown, random parameters $\theta = (a_0, a_x, a_z, b_0, b_x, b_{\tilde{x}})$. Theorem 7.4 shows that the joint distribution of θ, $G(\theta)$, is identified using local variation in y, $\tilde{x}$, x and z. Of course the linearity in (5) is just an example; Theorem 7.4 identifies a joint distribution over functions in a nonparametric function space.

The system (5) allows more general economic behavior than has previously been shown to be identified in the literature. In common with much of the literature, the response to $\tilde{x}$ is heterogeneous, as $b_{\tilde{x}}$ is a random coefficient. However, here the response to the instrument, a_z, is also a random coefficient. In contrast with the assumptions made in the literature on the local average treatment effect (LATE, see Imbens and Angrist (1994)) and some selection models (Vytlacil, 2002), some agents may have $a_z > 0$ and respond positively to the instrument, and other agents may have $a_z < 0$ and respond negatively to the instrument.[24] Further, the response to the instrument may be correlated with the response to the treatment. The joint distribution $G(\theta)$ may be such that those agents with the most to the gain from the treatment (a high marginal effect $b_{\tilde{x}}$) tend to have a high a_z. For a given z, this model allows agents to sort into an intensity of treatment $\tilde{x}$ based on the expected gains from treatment, $b_{\tilde{x}}$.

Consider an example. Firms differ in both their input demand functions (the first stage) and their production functions. Let y be the log output of a firm, x the age of the firm (which is independent of θ), $\tilde{x}$ the log number of workers hired by the firm (an endogenous choice variable), and z the price of labor. In this example, variation in input costs allow identification of the distribution of production functions in some industry. This framework is general. First, firms vary in how labor inputs affect outputs: the labor input elasticity $b_{\tilde{x}}$ is heterogeneous. Second, firms with higher labor elasticities may have higher input demand elasticities: $\text{Corr}(a_z, b_z) > 0$. Third, there is no monotonicity in a_z, some firms may have $a_z < 0$. Say the price of labor goes up everywhere and workers are laid off at some firms. Then, due to a general equilibrium effect, some firms might actually increase their labor inputs. Identification of $G(\theta)$ allows the identification of

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

the joint distribution of a_z and b_z as well as of the other coefficients.

7.3 Endogenous Regressors in Multinomial Choice

We now consider the endogeneity problem that arises in multinomial choice. Recalling the discussion of the multinomial choice model in Section 6.1, an endogeneity problem arises when an agent's preferences as captured by the utility function u are not independent of some elements of the agent's choice set (v, w). Such endogeneity could arise if, for example, the choice set $x = (v, w)$ that an agent faces is partly "designed" on the basis of information related to its type or preferences u. A classic example of this source of endogeneity arises in a principal-agent relationship, in which the principal designs the menu of contracts (v, w) facing the agent using information that is correlated with the agent's type u but that is not observable by the econometrician. 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.[25]

In this section, we show how to solve the endogeneity problem posed by endogenous product characteristics in multinomial choice by way of a triangular system of equations that follows much the same logic as endogenous regressors in nonadditive random functions. Essentially, the triangular system jointly models the decisions of both the principal and the agent, and uses exogenous variation in the characteristics of the principal-agent relationships to achieve identification. Recall the notation from Section 6.1 in which an agent is described by a utility function $u : \mathbb{R}^K \to \mathbb{R}^J$, and given $v \in \mathbb{R}^K$ and $w \in \mathbb{R}^J$, the agent has utility for choice j given by $u^j(v) + w_j$. Following the notation of Section 7.1, we let the first M elements of the vector of choice characteristics v be potentially endogenous, and denote these elements by $\tilde{v} \in \mathbb{R}^M$ and the remaining exogenous elements by $v \in \mathbb{R}^N$ where $N = K - M$. We refer to these endogenous elements $\tilde{v}$ as the principal's "prices" as they are strategically set by the principal.

To handle the problem, we introduce a vector of instruments $z = (z_1, \ldots, z_M) \in Z \subseteq \mathbb{R}^M$ that are stochastically independent of preferences u. In addition, the instruments are capable of shifting the endogenous choice characteristics through the principal's "pricing;; or IV equation $\tilde{v} = h(v, z)$ for $z \in Z$, $v \in V$, and $h : X \times V \to \mathbb{R}^M$.[26] Thus a type corresponds to a pair of functions (u, h) consisting of a vector valued utility function and an IV equation. The model is such that for any economic environment $x = (v, w, z)$, the response $f((v, w, z), (u, h))$ consists of the principal's

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

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

choice of prices $\tilde{v}$ and the agent's choice of product j that are linked through the recursive system

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

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

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

Assumption 7.5. *Assume that (v, w, z) has support equal to the product set $\mathcal{X} = V \times \mathbb{R}^J \times Z$, where $V \subseteq \mathbb{R}^N$ and $Z \subseteq \mathbb{R}^M$ are non-empty rectangles.*[27]

We assume that utility functions lie in a set satisfying the SNTP and the pricing/IV equations lie in a set satisfying the WNTP and satisfy a similar instrumental variable assumption as used in the previous section.

Assumption 7.6. *The type space Θ is a subset of $\mathcal{S}^{K,J}_{\mathbb{R}^K} \times \mathcal{W}^{K,M}_{V \times Z}$ such that the following conditions hold: (i) For any type $(u, h) \in \Theta$, the derivative $D_z h(v, z)$ of a type's IV equation with respect to the instruments everywhere exists in the interior of $V \times Z$, and is continuous in $(v, z) \in X \times Z$; and (ii) For any type $(u, h) \in \Theta$, and for any interior $v \in V$, the derivative $D_z h(v, z)$ with respect to z has full rank J for almost all (in the sense of Lebesgue measure) $z \in Z$.*

[27] Thus we let the special regressor have full support and no longer require the pure characteristics assumption. We could alternatively impose the pure characteristics assumptions and instead let the special regressor have arbitrarily small support.

Finally the stochastic independence between the exogenous regressors and the type

Assumption 7.7. *The instruments and exogenous regressors $(v, w, z) \in \mathcal{X}$ are statistically independent of the type $(u, h) \in \Theta$.*

Our main result is that the endogenous multinomial choice model is reducible and hence identifiable.

Theorem 7.8. *Under assumptions 6.3, 7.5, 7.6, and 7.7, the distribution of $(u, h) \in \Theta$ in the multinomial choice model with endogenous regressors is identified with respect to $\tilde{\mathcal{G}}$, the class of finite distributions.*

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

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

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

The proof for showing the existence of such a singleton I-set exactly follows the proof of Theorem 7.4, except with a relabelling of the relevant terms. In particular, $u, v, \tilde{v}$ play the role of $g, x, \tilde{x}$, respectively, from this previous proof, while h and z play the same role in both contexts. Replacing these terms (and adjusting the relevant dimensions of the model), the proof can be followed exactly until the end of case 2. Instead of picking an arbitrary point $u^* \in \mathbb{R}^J$ from the set of values $\{u(h_1(v', z'), v') \mid u \in T_3\}$, we instead pick a minimal element, which by construction is attained at a unique $u \in T_3$, which we denote u_1. Then setting the special regressors w to $w^* = -u_1$, we have that $I^T_{(0,h_1(v',z')),(v',w^*,z')}$ is a singleton, consisting of only the type $(u_1, h_1) \in T$. $\square$

7.4 Literature review on endogenous regressors

Our results on endogenous regressors are particularly notable. For example, Chesher (2003) studies the nonparametric identification of a triangular system of equations where the functions in the system are non-random: the same for all types. Heterogeneity enters only through scalar error terms in each equation, and those error terms are assumed to enter the non-random functions monotonically. We allow each type to have its own function and we impose no monotonicity assumptions about how unobservables relate to outcome variables and endogenous regressors. We also do not impose monotonicity assumptions on how the instruments affect endogenous regressors, which are common in the literature on treatment effects (Imbens and Angrist, 1994; Vytlacil, 2002). Newey and Powell (2003) use a mean independence assumption in a model where heterogeneity

enters the outcome equation as only an additive error, instead of a random function. Imbens and Newey (2008) study a system (g, h) like ours, except that the heterogeneity in the IV equation h is restricted to be a scalar. We allow h to be a random, nonadditive function. Further, Imbens and Newey require the scalar disturbance to enter h strictly monotonically. Imbens and Newey also define to the object of interest to be what they describe as a quantile structural function. We show the full identification of all aspects of our model, namely the joint distribution of the heterogeneous functions (g, h). There are many other approaches in the nonparametric instrumental variables literature (see the above papers for more references); we know of no others that identify a distribution over systems of functions.

Hoderlein, Klemelä and Mammen (2008) examine a linear triangular system such as (5), except that the coefficients a_0, a_x, a_z from the first stage are homogeneous. Only the parameters in the outcome equation are heterogeneous. Their approach relies critically on linearity, while we identify a nonparametric distribution on a nonparametric class of functions.

As discussed above, Berry and Haile (2008) and our paper simultaneously developed approaches to identifying the distribution of heterogeneity in multinomial choice models. An additional distinction is that 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 realizations within a market and variation across markets to address an endogeneity problem. One could replace their step where they trace out utility values with our Theorem 6.5.

8 Conclusions

There exist few nonparametric identification theorems for the distribution of heterogeneity in many economic models estimated every day in applied microeconomics. We introduce a property of economic models, known as separability, that is a sufficient condition for identification of the distribution of heterogeneity.

Our first application of separability is to identifying a distribution of nonadditive random functions. While others have explored distribution of random coefficients in the linear regression model or allowed for other aspects of unobserved heterogeneity, we are the first to work in the generality of identifying a distribution over a space of heterogeneous functions. We also explore identification of the distribution of marginal effects. The latter result does not rely on either the strong or weak no-ties properties.

In terms of multinomial choice, relative to the literature we have a least seven 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 Cramér and Wold theorem,

3) we identify the joint distribution of product-specific utility *functions* for all choices rather than just utility *values* conditional on v, 4) we are nonparametric on the subutility function $u^j(v)$ for choice j, 5) we allow for endogenous characteristics such as prices, 6) we show how to analyze multiple purchases when some goods can be complements and preferences may be correlated across multiple products, and 7) we show that we do not need large support if demand is given by the pure characteristics model.

Endogenous regressors are important in empirical work using observational data. We 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. We generalize 2SLS in important ways. First, all parameters can be random, with an unrestricted joint distribution. Thus, some agents may respond positively to the instrument and others may respond negatively. Further, the response to the instrument may be correlated with the response to the endogenous regressors: those with more to gain from treatment may adopt more intense treatments. Of course, we identify the first and second-stages of the triangular system nonparametrically: each agent is characterized by a pair of heterogeneous functions.

Our identification strategy, while not constructive, is compatible with the linear regression estimator of Bajari, Fox, Kim and Ryan (2009a) for the distribution of unobserved heterogeneity. The estimator has been proved to be a consistent estimator in the space of distribution functions for the unknown distribution G under a potential ill-posed inverse problem. Our discussion of identification complements the discussion of consistency (assuming identification) in Bajari et al.

A Identification With Positive Probability

Consider the model (1). To show the consistency of a nonparametric estimator for the distribution of heterogeneity, 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 $X^\star \subseteq \mathcal{X}$ with *positive probability* such that for all $x \in X^\star$, $\Pr_{G^0}(A \mid x) \neq \Pr_{G^1}(A \mid x)$ for some fixed $A \subseteq \mathcal{Y}$. We can this strong definition of identification "identification with positive probability."

As we now show, the existence of such a set positive measure $X^\star$ follows can follow easily from the existence of a pair $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$ for which $\Pr_{G^0}(A \mid x) \neq \Pr_{G^1}(A \mid x)$, as ensured by separability. In particular, we show that from the existence of such an experiment (y, x), we can find a small open ball $X^\star$ about x.

Lemma A.1. *Identification implies identification with positive probability if for any finite set of types $T \subset \Theta$, and for any $I^T_{A,x}$, there exists some small neighborhood $X^\star \subset \mathcal{X}$ containing x, where $z \in X^\star$ implies $\exists\, A_z$ such that $I^T_{A_z,z} = I^T_{A,x}$.*

Proof. We can always define G^0 and G^1 to assign probabilities to the same set of finite types $T = \{\theta_1, \ldots, \theta_n\} \subset \Theta$ 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^0_θ and p^1_θ. Let $A \subseteq \mathcal{Y}$ and $x \in \mathcal{X}$ be the experiment that distinguishes G^0 and G^1. Then we have

$$\sum_{\theta \in I^T_{A,x}} p^0_\theta \neq \sum_{\theta \in I^T_{A,x}} p^1_\theta.$$

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

$$\sum_{\theta \in I^T_{A_z,z}} p^0_\theta \neq \sum_{\theta \in I^T_{A_z,z}} p^1_\theta,$$

as $I^T_{A_z,z} = I^T_{A,x}$. □

We next consider two examples of applying this lemma.

Theorem A.2. *Under Assumptions 4.1, 4.2, and 4.3, the distribution of nonadditive random functions is identified with positive probability in the class $\tilde{\mathcal{G}}$ of finite distributions.*

Proof. Consider $I^T_{y,x} = \{g \in T \mid g(x) = y\}$. Let there be a singleton I-set. By continuity, changes in x in a small open set will not change the I-set $I^T_{\tilde{g}(x),x}$ indexed by a particular $\tilde{g} \in T$. Apply Lemma A.1. □

Theorem A.3. *Under assumptions 6.1, 6.2, 6.3, and 6.4, the distribution of utility functions in the multinomial choice model is identified with positive probability (in $x = (v, w)$) in the class $\tilde{G}$ of finite mixtures.*

Proof. Consider the I-set with respect to the outside good

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

There exists a singleton I-set $\{u_i\}$ by the argument in the proof of Theorem 6.5. Next, varying w to be smaller (more negative) in some small open set will cause type u_i to continue to pick the outside good. As w is varied, by a continuity argument v can be varied to preserve $I^T_{0,v,w} = \{u_i\}$. Exploiting the product support $V \times W \times \ldots \times W$, we can apply Lemma A.1. □

B Identification of Countable Distributions of Nonadditive Random Functions

In the main body of the paper, we alluded to the fact that identification with respect to countable rather than finite distributions (and hence showing countable separability rather than finite separability of the model) would rely on strengthening the no ties properties to extend to a countable subset of functions rather than a finite subset. We state the appropriate generalization here and show that the property immediately translates into a theorem that the set of nonadditive random functions satisfies countable separability. Recalling the functional space notation from Section 3.1, we can state the following definition.

Definition B.1. A subset $\mathcal{F}^{k,m} \subseteq \mathcal{C}^{k,m}$ satisfies the countable weak no ties property (countable WNTP) if for any countable subset $\{g_1, \dots\} \subset \mathcal{F}^{k,m}$ there exists $x \in \mathcal{X}$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \dots\}$.

Theorem B.2. *If the type space Θ satisfies the countable WNTP, then the non-additive random functions model (see Section 4) is identified with respect to $\mathcal{G}$, the set of countable distributions.*

Proof. We show countable separability of the model. Let $T = \{g_1, \dots\} \subset \Theta$ denote a countable set of types. As Θ satisfies the countable WNTP, then there exists $x \in \mathcal{X}$ such that $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \dots\}$, which we denote as x^*. Let y^* be any element from the set of values $\{g(x^*) \mid g \in T\}$, which is attained at say g_1, i.e., $g_1(x^*) = y^*$. Then by construction the I-set $I_{\{y^*\},x^*}$ consists of a single element, namely g_1. □

It is a natural question to ask whether any relevant function spaces $\mathcal{F}^{k,m} \subset \mathcal{C}^{k,m}$ satisfy the countable WNTP. A result by Reny (2008), in a followup to this paper, shows that the space $\mathcal{A}^{k,m}$ of vector valued real analytic functions (which recall contains the space of vector valued polynomials) satisfies the countable WNTP. We leave further demonstrations of the applicability of the countable WNTP to future work. By the continuity of the g functions, the countable WNTP can be restricted to apply when the space of covariates $\mathcal{X}$ is equal to the appropriate dimensional rational space. Thus, we can gain identification of a countable distribution of nonadditive functions g by using a countable (the rationals are countable) data generating process for the observables x. Observables and unobservables are treated symmetrically.

C Separability When Distributions Admit a Density Function

We wish to show that an analog to the concept of separability can be extended to models where the distribution $G(\theta)$ is required to admit a continuous density $\mu(\theta)$ over Θ. This is a non-nested class

to the class of countable (or finite) mixtures that we study. While we do not believe the extension is essential for practical applications in economics, it is of some theoretical interest because it shows that the ideas behind separability can be extended to other types of distributions. As before, let the economic model be $\mathcal{M}$. For any $T \subset \Theta$, the I-set $I_{y,x}^T = \{\theta \in T \mid f(x,\theta) \leq y\}$ may no longer be a finite or a countable set of points. In certain well-behaved models, we may imagine $I_{y,x}^T$ to be a subset of T with a non-empty interior. It is this non-empty interior that will make verifying the property that there exists a singleton I-set $I_{y,x}^T$ difficult. Instead, we propose another notion of separability that may be satisfied in some simple models.

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. All partitions of Θ are included in $\mathcal{A}$. An economic model $\mathcal{M}$ is **continuously separable** 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 $(y,x) \in \mathcal{Y} \times X$ and $i \in C_T$ for which $I_{y,x}^T \subseteq U_i$ for some open, connected $U_i \subset T$. This property must hold for every $T \in \mathcal{A}$.

Theorem C.1. *If the model $\mathcal{M}$ is continuously separable, then it is identified within the class of distributions with continuous densities over Θ.*

Proof. Suppose that continuous separability holds but that the model is not identified. Then, there exist two continuous densities $\mu_0(\theta)$, the truth, and $\mu_1(\theta)$ that both give the same distribution $F(y \mid x) = F_0(y \mid x) = F_1(y \mid x)$ for the data. Let $\pi(\theta) = \mu_0(\theta) - \mu_1(\theta)$. The function $\pi(\theta)$ is continuous because $\mu_0(\theta)$ and $\mu_1(\theta)$ are. Then

$$F_0(y \mid x) - F_1(y \mid x) = \int_\Theta \pi(\theta) 1[f(x,\theta) \leq y] \, d\theta = 0.$$

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,

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

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 continuous separability, there exists $(y,x) \in \mathcal{Y} \times X$ and $i \in C_T$ for which $I_{y,x}^T \subseteq U_i$ for some open, connected $U_i \subset T$. By the disjoint supports, either

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

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

so we have identification. □

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

Recall that we can use Zorn's lemma to show that there are maximal sets that satisfy the WNTP and the SNTP, respectively. However, we cannot otherwise describe those sets. One subset of the maximal sets for both the WNTP and the SNTP is the space of all real analytic functions. This appendix shows that the space of all real analytic functions satisfies both the WNTP and the SNTP.

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

Real analytic functions must be infinitely differentiable.

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

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

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

Proof. Consider any finite set of vector valued real analytic functions $\{g_1, \ldots, g_n\} \subset \mathcal{A}^{k,m}$. We show by induction on n that the property holds for any finite number of elements n. The base case $n = 2$ holds by the above property of vector valued real analytic functions (for any open set $U \subseteq \mathcal{X}$, take any non-empty ball within U, which is connected, and apply the pairwise tie breaking property to this ball). Assume that the proposition holds for $n - 1$, and consider $\{g_1, \ldots, g_n\}$ and an open set $U \subseteq \mathcal{X}$, which without loss we can take to be an open ball (U contains such a ball, and balls are connected). By the induction hypothesis, there exists a point $x \in U$ such that $g_i(x) \neq g_j(x)$ for any $g_i \neq g_j$ and $i, j \in \{1, \ldots, (n-1)\}$. By the fact each g_i is continuous and the set of functions is finite, these inequalities are preserved in a small open ball $B_1 \subseteq U$ around x. Now consider the function g_n, and observe that by the pairwise tie breaking property, there exists

an $x_1 \in B_1$ such that $g_n(x_1) \neq g_1(x_1)$. Furthermore, by continuity, this inequality is preserved in a small ball $B_2 \subseteq B_1$ containing x_1. Now repeat the argument, except comparing g_n with g_2, producing the a ball $B_3 \subseteq B_2$, etc. At the end of the process, a non-empty ball $B_n \subseteq B$ is produced for which any $x \in B_n$ satisfies the definition of the SNTP, i.e., $x \in B_n$ implies $g_i(x) \neq g_j(x)$ for any distinct g_i and g_j in $\{g_1, \ldots, g_n\}$. □

E Proof of Lemma 5.1

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

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

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

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

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

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

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

References

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

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," July 2009. University of Minnesota working paper.

_ , _ , _ , and _ , "The Random Coefficients Logit Model is Identified," April 2009. NBER working paper.

Barbe, Philippe, "Statistical analysis of mixtures and the empirical probability measure," *Acta Appl. Math*, 1998, *50* (3), 253–340.

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

Berry, Steven and Ariel Pakes, "The Pure Characteristics Demand Model," *International Economic Review*, 2007, *48* (4), 1193–1225.

Berry, Steven T. and Philip A. Haile, "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," 2008. Yale University 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," *Journal of Business and Economic Statistics*, 2009.

Burda, Martin, Matthew Harding, and Jerry Hausman, "A Bayesian Mixed Logit-Profit Model for Multinomial Choice," April 2008. Stanford University working paper.

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

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

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

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

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

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.

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

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

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

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

Heckman, James J. 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.

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

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

_ , Jussi Klemelä, and Enno Mammen, "Analyzing the Random Coefficient Model Nonparametrically," June 2008. Brown University working paper.

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

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

_ **and Whitney K. Newey**, "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," June 2008. Harvard University 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. University of Connecticut working paper.

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

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

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

_ , "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 2003, *71* (5), 1339–1375.

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

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

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

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

Reny, Philip, "An Identification Result Based on work by Gandhi and Fox," November 2008. University of Chicago 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.

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.

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

Yakowitz, Sidney J. and J. Sprangins, "On the Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 1968, *39*, 209–214.