NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION IN MODELS OF MULTINOMIAL CHOICE

(PRELIMINARY AND INCOMPLETE)

ERIC M. BECKER

YALE UNIVERSITY

Abstract. I consider identification and consistent estimation in models of multinomial choice. I discuss a new and very general identification theorem that applies to multinomial choice models with an additive independent logistically distributed error. I also offer two new consistency results for nonparametric maximum likelihood estimation in the linear utility random coefficients multinomial choice models.

Date: Sept 24, 2009.
Thanks to: Yuichi Kitamura, Xiaohong Chen, Peter C.B. Phillips, Phil Haile, Steve Berry, Zhipeng Liao, and Xiaoxia Shi for their support and feedback.

1. Introduction

Models of multinomial choice have been used in various applied fields including IO, marketing, city planning, political science, etc., to analyze and predict the behavior of a group utility maximizers. Specifically, these utility maximizers are decision makers in a population who don't interact and who are each faced with a (1) finite menu of options. It is typical to assume that (2) utility is linear and that (3) the functional form of utilities is known up to a finite dimensional parameter. Utilities are allowed vary from person to person but (4) the randomness enters utility in a specified way. Finally, (5) the distribution of randomness in these models is assumed to take a parametric form.

One could think about generalizing along any one of the points (1)-(5) above. This is probably the ultimate goal. In this paper I will talk only about removing assumption 5. I will give results about identification and consistency in two classes of multinomial choice models:

(1) the class of 'logit' multinomial choice models.

(2) the class of 'distribution free' multinomial choice models.

To make the easier to digest I will conform to the following notational scheme. Bold face type will be used for vectors, cursive will be used to denote collections. I will for the most part use capitals when referring to random variables and lower case letters when referring to constants. I define R^w to be euclidean space of dimension w and cR^w will be some compactification of R^w (possibly a different compactification in different places). Also, rR^w will denote the remainder $cR^w \setminus R^w$.

2. Related Literature

There have been a number of papers written on estimation in the binomial choice model. Gautier and Kitamura (2009) discuss estimation of an L2 density in the pure linear utility random coefficients model. They do not use npml, but it is interesting to note that they require a similar normalization to achieve identification and to proceed with consistent estimation. Ichimura and Thompson (1999) have

a paper about nonparametric maximum likelihood estimation in the linear utility binomial choice model. However, there is an error in the proof of their consistency theorem. I show how to correct this error below. There are differences in the way utilities are normalized that will have ramifications for estimation.

Newey and Powell (2003) study nonparametric IV estimation. Their results can be applied to this model, which would require I use the compactification device cited therein. I chose not to go that route because the spaces under use there are not probabilistic in nature.

Ichimura and Thompson (1999) discuss the npml estimation of a mixing distribution in the binomial choice, linear utility, distribution free model. Their paper turns out to have an error that invalidates the consistency result stated therein. Part of the contribution of this work is to show how their result can be fixed. Indeed the corollary below that applies to the linear utility, distribution free example is analogous to the model in that paper. Although the assumptions needed to achieve consistency are similar, a different normalization is used in this paper that has implications if one were to apply the estimator in practice.

The literature on nonparametric maximum likelihood in the context of mixing problems is extensive. Methods for consistently estimating a mixing distribution (and sometime simultaneously a parametric finite dimensional parameter) have been studied by Wald (1949), Kiefer and Wolfowitz (1956), Pfanzagl (1988), van der Vaart (1988), and many others. The typical consistency theorem in these works require a kernel (see below) that is smooth and that vanishes at infinity. I will talk more about this below. These techniques have to be adapted in order to make them useful in the mulitnomial choice framework.

3. Mixture Models

3.1. **A motivating example.** To motivate definitions, consider the following scenario. A Yale econometrics student, call him student i, has booked a flight home

for the holidays through Laguardia International airport. He has to arrange transportation to the airport and can choose choose one of the following options: (1) Drive to the airport and park in the lot; (2) Take the CT limo direct to the airport; (3) Take the train to grand central and the rail to Laguardia; or (0) Do something else (cancel trip, walk to the airport, etc.).

When making his decision the traveler takes into consideration: total cost, congestion, commute time, amount of walking required, weight of baggage, and 'mood,' a characteristic that is only observable to student i. For simplicity, assume that the costs of each option, commute time, baggage weight, and congestion are exogenously determined. Student i's utility for each option takes the following form (set $\overline{X} \equiv X - \mathbf{E}(X)$).

$$\begin{aligned} u_{i,j} =& -10(\overline{\text{cost}})_{i,j} - (\overline{\text{cmmt time}})_{i,j} \\ & - 2(\text{bag wght})_i(\overline{\text{walkreqrd}})_{i,j} - 3(\overline{\text{cngstn}})_{i,j} + (\text{mood})_{i,j} \\ u_{i,0} =& 0 \end{aligned}$$

Notice that I have demeaned the values of the observables variables. This is necessary because identification in this model will require that covariates vary within a neighborhood of the origin. One may assume the averages are absorbed into the $(\text{mood})_j$ term. Student i chooses option j if his utility for that option is the greatest. A 'model' describing student i's choice is:

$$C_{i,j} = 1\{u_{i,j} = \max_{k=0,1,2,3} u_{i,k}\}$$

We may believe the form of utility above does a good job of describing the behavior of all students faced with such a decision, but that students differ in the weighting they give to each characteristic. Then we could use the collection $\mathcal{C} = \{\mathbf{C}(\mathbf{x}, \mathbf{z}; \mathbf{mood}, \boldsymbol{\zeta}) \ : \ (\mathbf{mood}, \boldsymbol{\zeta}) \in H\}$ to study the behavior of that group. Here $\mathbf{z} = ((\text{cngstn})_j)_{j=1}^3$, $\mathbf{x}$ is the vector of other observables, and $\boldsymbol{\zeta}$ indexes admissible weighting schemes. In the example above $\boldsymbol{\zeta} = \{-10, -1, -2, -3\}$.

If the distribution of $(\mathbf{mood}, \boldsymbol{\zeta})$ in the population is given by G, then the behavior of a randomly drawn individual can be described by

$$\mathbf{s}(\mathbf{x}, \mathbf{z}; G) = \int \mathbf{C}(\mathbf{x}, \mathbf{z}; \mathbf{mood}, \boldsymbol{\zeta}) dG(\mathbf{mood}, \boldsymbol{\zeta}).$$

Finally, if the distribution is only known to lie in a family $\mathcal{P}$ then we could use the collection $\mathcal{Y} = \{\mathbf{s}(\mathbf{x}, \mathbf{z}; G) : G \in \mathcal{P}\}$ to analyze the behavior of a random sample of observations in the population.

These definitions are meant to motivate the general definition of a mixture model that follows. This is the framework that will be used to analyze data in a multinomial choice setting.

3.2. **The mixture model and assumptions.** I will do my best to follow Bickel et al. in my notation. This is not entirely possible because the class of mixing models studied there do not encompass the models under study here. This is in part due to the fact that the distribution of covariates $(\mathbf{X}_t, \mathbf{Z}_t)$, a nuisance parameter in this context, does not enter in the reduced form (it is not estimated).

Begin with the model $\mathcal{C} = \{\mathbf{C}(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) : \theta \in \Theta, \boldsymbol{\eta} \in H\}$. In this paper $\mathcal{C}$ will typically be a collection of indicator functions taking as input a vector $(\mathbf{x}, \mathbf{z})$, with $\mathbf{x} \in R^{Jd_x}$ and $\mathbf{z} \in R^J$. Here J is the number of products available for purchase, so that each consumer has $J + 1$ options in their choice set. I assume, for each $(\mathbf{x}, \mathbf{z})$, for each $\theta \in \Theta$, the functions $\mathbf{C}(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) \in \mathcal{C}$ are measurable with respect to the borel sigma field on H, denoted $\mathcal{B}_H$.

With this assumption I can define $\mathcal{Y} = \{\mathbf{s}(\mathbf{x}, \mathbf{z}; \theta, \mathrm{G}) : \theta \in \Theta, \mathrm{G} \in \mathcal{P}\}$, where $\mathcal{P}$ is some large space of probability measures on $(H, \mathcal{B}_H)$. For each fixed value of $(\mathbf{x}, \mathbf{z})$

$$\mathbf{s}(\mathbf{x}, \mathbf{z}; \theta, \mathrm{G}) = \int_H \mathbf{C}(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) d\mathrm{G}(\boldsymbol{\eta}).$$

Let $(\mathbf{X}_t, \mathbf{Z}_t)$ be an observable r. vector with distribution function $F_{(\mathbf{x},\mathbf{z})}$. In most places I will assume $F_{(\mathbf{x},\mathbf{z})}$ has a density with respect to lesbegue measure, which I will denote $f_{(\mathbf{x},\mathbf{z})}$. For most of this paper I will assume $(\mathbf{X}_t, \mathbf{Z}_t)$ is independent of $\boldsymbol{\eta}$.

Let $\mathbf{Y}_t = (Y_{0,t}, Y_{1,t}, ..., Y_{J,t})$ denote the r. vector where $Y_{j,t}$ takes the value 1 if and only if product j is chosen. If the true parameter is (θ, G) then $\mathbf{Y}_t$ is related to $(\mathbf{X}_t, \mathbf{Z}_t)$ through $\mathbf{s}$:

$$\Pr\{Y_{j,t} = 1 \mid (\mathbf{X}_t, \mathbf{Z}_t) = (\mathbf{x}, \mathbf{z})\} = s_j(\mathbf{x}, \mathbf{z}; \theta, \mathrm{G}).$$

Then the joint density of $(\mathbf{Y}_t, \mathbf{X}_t, \mathbf{Z}_t)$ at $(\mathbf{y}, \mathbf{x}, \mathbf{z})$ is given by

$$\mathbf{s}(\mathbf{x}, \mathbf{z}; \theta, \mathrm{G}) f_{(\mathbf{x},\mathbf{z})}(\mathbf{x}, \mathbf{z}). \tag{1}$$

For easy reference I summarize the assumptions leading to this reduced form below.

(M1) For each $(\mathbf{x}, \mathbf{z})$, for each $\theta \in \Theta$, the functions $\mathbf{C}(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) \in \mathcal{C}$ are measurable with respect to the borel sigma field on H.

(M2) The r. vector $(\mathbf{Y}_t, \mathbf{X}_t, \mathbf{Z}_t)$ is i.i.d. across t, the r. vector $(\mathbf{X}_t, \mathbf{Z}_t)$ is independent of $\boldsymbol{\eta}$. Also, $F_{(\mathbf{x},\mathbf{z})}$ has a density with respect to lebesgue measure on $R^{J(d_x+1)}$. The statistical relationship between $\mathbf{Y}_t$ and $(\mathbf{X}_t, \mathbf{Z}_t)$ is given in eqn. (1).

As mentioned previously, I will ignore the nuisance density $f_{(\mathbf{x},\mathbf{z})}$. One may need to bring it back into the model to talk about efficiency. I will next talk about the class of mixing problems relevant to this paper

3.3. **Multinomial choice, criterion, and estimators.** This research considers the class of multinomial choice models. In this category $\mathcal{C}$ will be a collection of indicator functions defined as follows. Let the utility derived from product j when the observable characteristics take the value $(\mathbf{x}, z) \in R^{d_x+1}$ and the incidental parameter is $\boldsymbol{\eta}$ be given by:

$$u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = V_j(\mathbf{x}_j, \theta, \boldsymbol{\eta}) + z_j$$

$$u_0 = 0$$

The identity $u_0 = 0$ is without loss of generality because only the difference between utilities will factor into consumers' decisions. In addition, a scale normalization is needed. In this paper I normalize the coefficient on z_j to be 1. To fully define $\mathbf{C}$ I will need to first define the space H. For now assume $\boldsymbol{\eta} \in R^{d_\eta}$ and set

$$C_j(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) \equiv 1\{u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = \max u_k(\mathbf{x}_k, z_k, \theta, \boldsymbol{\eta})\}, \tag{2}$$

with $(\mathbf{x}, \mathbf{z})$ arbitrary. I have assumed $\mathbf{u}$ is known up to the parameter θ. Of course I will need to verify condition (M1) in each example. Also notice that for any $\boldsymbol{\eta}$ such that $u_j = u_k$ we have $\sum_j C_j(\mathbf{x}, \mathbf{z}; \theta, \boldsymbol{\eta}) > 1$. This is actually a problem, and I will need to address it if I expect the criterion to obtain a maximum at (θ_0, G_0) (see below).

With definitions as in the previous section, and assuming (M1) holds, I can define the following two criterion functions for M-estimation in models of multinomial choice. In this paper I consider consistent estimation only in problems without a finite dimensional parameter. The first estimator is the maximum likelihood estimator, the criterion is

$$Q^{FML}(G) = \mathbf{E}\left[\sum_{j=0}^{J} Y_{j,t} \log(s_j(\mathbf{X}_t, \mathbf{Z}_t; G))\right].$$

The empirical criterion is the log-likelihood, which is given by

$$\hat{Q}_T^{FML}(G) = \sum_{t=1}^{T}\sum_{j=0}^{J} Y_{j,t} \log(s_j(\mathbf{X}_t, \mathbf{Z}_t; G)).$$

So the nonparametric full maximum likelihood (npfml) estimator for G is given by

$$\hat{G}_T^{FML} = \text{argmax}_{G \in \mathcal{P}} \hat{Q}_T^{FML}(G).$$

The second class of estimators are based on a criterion involving the conditional likelihood for the outside option only. That criterion function is

$$Q^{PML}(G) = \mathbf{E}\left[Y_{0,t} \log(s_0(\mathbf{X}_t, \mathbf{Z}_t; G)) + (1 - Y_{0,t}) \log(1 - s_0(\mathbf{X}_t, \mathbf{Z}_t; G))\right].$$

The empirical criterion is the partial log-likelihood, which is given by

$$\hat{Q}_T^{PML}(G) = \sum_{t=1}^{T} \Big[Y_{0,t} \log(s_0(\mathbf{X}_t, \mathbf{Z}_t; G)) + (1 - Y_{0,t}) \log(1 - s_0(\mathbf{X}_t, \mathbf{Z}_t; G)) \Big].$$

So the nonparametric partial maximum likelihood (nppml) estimator for G is given by

$$\hat{G}_T^{PML} = \text{argmax}_{G \in \mathcal{P}} \hat{Q}_T^{PML}(G).$$

In the next subsection I will talk about some examples that fit into the framework discussed above.

3.4. **Some examples.** The following examples should help to clarify the concepts discussed above. In examples 3 and 4 below the functions C_j for which $\boldsymbol{\zeta} \in rR^{d_\zeta}$ have yet to be defined. Also, assumption (M1) has to be verified in each example.

VERIFICATION OF MEASURABILITY?

The first two examples will be used in the section on identification. The second two examples will be used to discuss estimation. Begin with

Example 1 (general utility distribution free)**.** In the sequel I will refer to this example as the distribution free, general utility model. The only incidental parameters are the unknown random coefficients. In order to be consistent I will always use $\boldsymbol{\zeta}$ to refer to the random coefficients (refer to the other examples below). I assume $\boldsymbol{\zeta}$ takes its values in the R^{d_ζ} so that $\mathrm{H} = R^{d_\zeta}$. Then $\boldsymbol{\zeta} \equiv \boldsymbol{\eta}$ and $V_j(\mathbf{x}_j, \theta, \boldsymbol{\eta}) = \overline{V}_j(\mathbf{x}_j, \theta, \boldsymbol{\zeta})$. Therefore, u_j is given by

$$u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = \overline{V}_j(\mathbf{x}_j, \theta, \boldsymbol{\zeta}) + z_j$$
$$u_0 = 0.$$

The collections $\mathcal{C}$ and $\mathcal{Y}$ are defined using (2) above and the definitions from section 3.2. ★

Example 2 (general utility logit)**.** I will refer to this example as the general utility logit model. In this model θ does not enter. The incidental parameters include the unknown random coefficients as well as an additive independent logistically distributed J-dimensional r. vector. Here $\boldsymbol{\eta} = (\boldsymbol{\zeta}', \varepsilon_1, ..., \varepsilon_j)'$. Again I will assume $\boldsymbol{\zeta}$ takes its values R^{d_ζ}. Thus $\mathrm{H} = R^{d_\zeta} \times R^J$.

Now $V_j(\mathbf{x}_j, \theta, \boldsymbol{\eta}) = \overline{V}_j(\mathbf{x}_j, \theta, \boldsymbol{\zeta}) + \varepsilon_j$. For now I assume the variance of the logistic error is known and $\mathrm{var}(\varepsilon_j) = \pi^2/3$. This is not without loss of generality, but see below. Therefore, u_j is given by

$$u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = \overline{V}_j(\mathbf{x}_j, \theta, \boldsymbol{\zeta}) + z_j + \varepsilon_j$$
$$u_0 = 0.$$

The collections $\mathcal{C}$ and $\mathcal{Y}$ are defined using (2) above and the definitions from section 3.2. ★

Example 3 (linear utility distribution free)**.** This example describes the distribution free, linear utility random coefficients model. In this model θ does not enter. The only incidental parameters are the unknown random coefficients. As in the examples above I use $\boldsymbol{\zeta}$ to refer to the random coefficients. I assume $\boldsymbol{\zeta}$ takes its values in some compactification cR^{d_ζ} of the space R^{d_ζ} so that $\mathrm{H} = cR^{d_\zeta}$ (see section 5.3 for details). Then $\boldsymbol{\zeta} \equiv \boldsymbol{\eta}$ and $V_j(\mathbf{x}_j, \theta, \boldsymbol{\eta}) = \mathbf{x}_j'\boldsymbol{\zeta}$. Thus

$$u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = \mathbf{x}_j'\boldsymbol{\zeta} + z_j$$
$$u_0 = 0.$$

The collections $\mathcal{C}$ and $\mathcal{Y}$ are defined using (2) above and the definitions from section 3.2. ★

Example 4 (linear utility logit)**.** I will refer to this example as the logit, linear utility model. In this model θ does not enter. The incidental parameters include the unknown random coefficients as well as an additive independent logit error. Here

$\boldsymbol{\eta} = (\boldsymbol{\zeta}, \varepsilon_1, ..., \varepsilon_j)$. Again I will assume $\boldsymbol{\zeta}$ takes its values in some compactification cR^{d_ζ} of the space R^{d_ζ}. Thus $\mathrm{H} = cR^{d_\zeta} \times R^J$.

Here $V_j(\mathbf{x}_j, \theta, \boldsymbol{\eta}) = \mathbf{x}_j'\boldsymbol{\zeta} + \varepsilon_j$. For now I assume the variance of the logistic error is known and $\mathrm{var}(\varepsilon_j) = \pi^2/3$. This is not without loss of generality, but see below. Therefore

$$u_j(\mathbf{x}_j, z_j, \theta, \boldsymbol{\eta}) = \mathbf{x}_j'\boldsymbol{\zeta} + z_j + \varepsilon_j$$

$$u_0 = 0.$$

The collections $\mathcal{C}$ and $\mathcal{Y}$ are defined using (2) above and the definitions from section 3.2. ★

This concludes the examples that will be discussed in this paper. In some cases it is not unreasonable to assume $\mathrm{var}(\varepsilon_j)$ is known. For example, it is often the case that the logit specification is used as a way to smooth the distribution free model, making estimation more tractable. In this case we can view the standard deviation on $\boldsymbol{\varepsilon}$ as a tuning parameter. To do this, simply place a fixed weight on z_j. The larger the weight the smaller the variance on ε_j. In the next section I talk about identification.

4. Identification

4.1. **Preliminaries.** In this section I will talk about identification in models of multinomial choice. I begin with a definition.

Definition. Call the equation defining $\mathbf{s}$ in terms of $\mathbf{x}, \mathbf{z}$ and the parameters θ, G a reduced form. I will say that θ_0, G_0 is identified in the set $\Theta, \mathcal{P}$ if no other parameter generates the same distribution of observables via the reduced form.

I will be primarily interested in identifying the image measure of $\boldsymbol{\zeta}$ under the mapping $\overline{\mathbf{V}}(\mathbf{x}; \theta, \cdot)$. To talk about this intelligently I need to be precise about what it means for a functional of the parameters to be identified. This motivates the following definition.

Definition. Let $\kappa(\theta, G)$ be a functional of the parameters. I will say that κ is identified in the set $\Theta, \mathcal{P}$ if for any two sets of parameters θ_1, G_1 and θ_2, G_2 generating the same distribution of observables we have that $\kappa(\theta_1, G_1) = \kappa(\theta_2, G_2)$.

MEASURABILITY HERE?

The rest of this section deals with results conditional on $\mathbf{X}_t$, so for now I will suppress $\mathbf{X}_t$ from the notation. As a first taste of identification in these models I will consider the general utility distribution free model (example 1).

4.2. **Identification in example 1.** If we assume that the conditional on $\mathbf{X}_t$ distribution of $\mathbf{Z}_t$ has full support in R^J then identification in this model follows easily from the observation that

$$\begin{aligned} s_0(\mathbf{Z}_t; \theta, G) &= \int 1\{\overline{V}_j(\theta, \boldsymbol{\zeta}) + Z_{j,t} \leq 0 \ \ \forall\ j > 0\} dG(\boldsymbol{\zeta}) \\ &= \int 1\{-\lambda_j \geq Z_{j,t}\ \forall\ j > 0\} d\overline{V}_j(\theta, G)(\boldsymbol{\lambda}) \\ &= \overline{V}_j(\theta, G)(\{\boldsymbol{\lambda} \leq -\mathbf{Z}_t\}), \end{aligned}$$

(see Berry and Haile (2008)). In the above display $\overline{V}_j(\theta, G)$ is the image measure of G under the mapping $\overline{V}_j(\theta, \cdot)$. Notice that $s_0(\mathbf{Z}_t; \theta, G)$ is the distribution of $Y_{0,t}$ conditional on $(\mathbf{X}_t, \mathbf{Z}_t)$. These observations lead to the following identification result.

Proposition 1. Assume (M1) and (M2) hold. Assume the the r. vector $\mathbf{Z}_t$ has full support in R^J conditional on $\mathbf{X}_t = \mathbf{x}$. Then $\overline{V}_j(\mathbf{x}; \theta, G)$ is identified in Exmaple 1 above.

Note: measurability restrictions on Θ and $\overline{V}_j(\theta, \boldsymbol{\zeta})$ are needed. The restriction on the support of $\mathbf{Z}_t$ is strong. I will show below that this later restriction can be weakened significantly in the general utility logit model (example 2).

In the section on consistency I will talk more about how results of this type can be extended to identification results on the parameters themselves. For now I will be content to talk about identification of functionals of this type.

4.3. **Identification in example 2, introduction.** I will now talk about identification in the general utility logit model (example 2). In this example the distribution of $\mathbf{Y}_t$ conditional on $(\mathbf{X}_t, \mathbf{Z}_t)$ can be written in a different form. This can be seen by integrating out the logistic errors. Indeed

$$\begin{aligned}\mathbf{s}(\mathbf{Z}_t;\theta,\mathrm{G}) &= \int 1\{\overline{V}_j(\theta,\boldsymbol{\zeta}) + Z_{j,t} + \varepsilon_j = \max \overline{V}_k(\theta,\boldsymbol{\zeta}) + Z_{k,t} + \varepsilon_k\} d(F_{\boldsymbol{\varepsilon}} \times G)((\boldsymbol{\eta}) \\ &= \int \mathbf{L}(\overline{\mathbf{V}}(\theta,\boldsymbol{\zeta}) + \mathbf{Z}_t) dG(\boldsymbol{\zeta}) \\ &= \int \mathbf{L}(\boldsymbol{\lambda} + \mathbf{Z}_t) d\overline{V}_j(\theta, G)(\boldsymbol{\lambda}).\end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\zeta}, \varepsilon_{i1}, ..., \varepsilon_j)$ and

$$\begin{aligned}\mathrm{L}_j(\overline{\mathbf{V}}(\theta,\boldsymbol{\zeta}) + \mathbf{Z}_t) &= \frac{\exp(\overline{V}_j(\theta,\boldsymbol{\zeta}) + Z_j)}{1 + \sum \exp(\overline{V}_k(\theta,\boldsymbol{\zeta}) + Z_k)} \qquad j \neq 0 \\ \mathrm{L}_0(\overline{\mathbf{V}}(\theta,\boldsymbol{\zeta}) + \mathbf{Z}_t) &= \frac{1}{1 + \sum \exp(\overline{V}_k(\theta,\boldsymbol{\zeta}) + Z_k)},\end{aligned}$$

see e.g., Traine (2009). To prove identification of $\overline{V}_j(\theta, G)$ I will first use derivatives of the market share function $s_0(\mathbf{x}, \mathbf{z}, \theta, G)$ to demonstrate that a certain collection of moments are identified. I then prove that the given collection is rich enough to uniquely determine $\overline{V}_j(\theta, G)$. In order to make this precise I will recall some definitions and preliminary results.

4.4. **Identification in example 2, preliminaries.** In order to make this precise I will recall some definitions and preliminary results.

Definition. A collection of functions $\mathcal{A}$ is called 'determining' or a 'determining class' for the collection of probability measures $\mathcal{P}$ when the following statement holds. If $G(f)$ is observable for all $f \in \mathcal{A}$ then G is identified in $\mathcal{P}$.

I will also need to use some basic concepts from function algebra.

Definition. An algebra A of functions $f : R^n \to R$ is a vector space of functions over the field of real numbers that is also closed under multiplication of functions. For any collection of functions $\mathcal{L}$ the algebraic closure, $A(\mathcal{L})$, of $\mathcal{L}$ is the closure of $\mathcal{L}$ under the operations of addition, multiplication, and scalar multiplication.

The following is a version of Stone's generalization of Weierstrass' approximation theorem.

Theorem 1. Let $\mathcal{A} \in C_0(R^w)$ be a subalgebra of continuous functions that (1) vanish at infinity such that (2) $\mathcal{A}$ separates points in R^w and (3) for each $\mathbf{x} \in R^w$ there exists an $f \in \mathcal{A}$ such that $f(\mathbf{x}) \neq 0$. Then $\mathcal{A}$ is dense in $C_0(R^w)$ under sup norm.

For a proof see Conway (1990). This leads to the following theorem about determining classes. For a proof see appendix A.

Theorem 2. Let $\mathcal{A}$ be a subalgebra of continuous functions that vanish at infinity satisfying conditions (1)-(3) of the previous theorem, then $\mathcal{A}$ is determining.

The following example may help to clarify how these concepts relate to identification.

Example 5. Let $\mathcal{P}([0,1])$ be the collection of probability measures on $([0,1], \mathcal{B})$. It is well known that the moments of a probability distribution are determining within the class $\mathcal{P}([0,1])$. We can use the above concepts to prove this fact. Since $\mathcal{P}([0,1]) \subset C_b([0,1])'$, where $C_b([0,1])'$ is the dual space of $C_b([0,1])$, we have $P_1 \neq P_2 \in \mathcal{P}([0,1])$ only if there exists an $f \in C_b([0,1])$ with $P_1(f) \neq P_2(f)$. By Weierstrass' original approximation theorem, the collection of polynomials is dense under sup norm in $C_b([0,1])$. Therefore $P_1 \neq P_2$ if and only if there is a polynomial p with $P_1(p) \neq P_2(p)$. But the collection of moments are identified if and only if the expectation of every polynomial is known and therefore $P_1 \neq P_2$ if and only if there is an m with $P_1(x^m) \neq P_2(x^m)$. ★

I will follow the format of this example to prove identification in general utility logit model.

4.5. **Identification in example 2, results.** I now proceed to state and prove a result about identification of $\overline{V}_j(\theta, G)$ in example 2. Define $\mathcal{L} = \{\mathrm{L}_j(\overline{\mathbf{V}}(\theta, \boldsymbol{\zeta}) + \mathbf{z}_0) : j \in J \cup \{0\}\}$ for a fixed $\mathbf{z}_0$. It turns out that $A(\mathcal{L})$ is determining for the collection of finite borel measures $\mathcal{M}(R^J)$. Indeed, the following proposition is true

Proposition 2. Assume (M1) and (M2) hold. Further assume conditional on $\mathbf{X}_t$ the support of $\mathbf{Z}_t$ contains an open set in R^J. Then $\overline{V}_j(\theta, G)$ is identified in example 2.

proof. I will first show that the collection $\{\overline{V}_j(\theta, G)(f) : f \in A(\mathcal{L})\}$ can be derived from $s_0(\mathbf{z}; \theta, G)$. This will imply that $\overline{V}_j(\theta, G)_1(f) \neq \overline{V}_j(\theta, G)_2(f)$ for some $f \in A(\mathcal{L})$ only if $s_0(\mathbf{z}; \theta, G_1) \neq s_0(\mathbf{z}; \theta, G_2)$. I prove this fact using induction. Notice that since expectation is linear, in order to derive $\{\overline{V}_j(\theta, G)(f) : f \in A(\mathcal{L})\}$, I only need to derive the moments of functions in the closure of $\mathcal{L}$ under multiplication.

Begin by noticing $\{\overline{V}_j(\theta, G)(f) : f \in \mathcal{L}\} = \{s_j(\mathbf{z}; \theta, G) : j \in J \cup \{0\}\}$, which shows that $\{\overline{V}_j(\theta, G)(f) : f \in \mathcal{L}\}$ are directly observable from the conditional on $(\mathbf{X}_t, \mathbf{Z}_t)$ distribution of $\mathbf{Y}_t$. This forms the base step of the induction argument. For the induction step, let $\boldsymbol{\alpha}$ be a multi-index and let $\mathbf{e}_j$ be a standard basis vector in R^{J+1}. First suppose $j \neq 0$ and suppose we have derived an expression for $E(\mathbf{L}^{\boldsymbol{\alpha}}$ in terms of $\{s_j(\mathbf{z}; \theta, G) : j \in J \cup \{0\}\}$. Notice

$$\begin{aligned} D_{z_j}\mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}}) &= \mathbf{E}(D_{z_j}(\mathbf{L}^{\boldsymbol{\alpha}})) \\ &= -(\alpha_0 + \alpha_1 + ... + \alpha_J)\mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}} L_j) + \alpha_j \mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}}), \end{aligned}$$

where the first equality follows by the dominated convergence theorem (see Dudley, 2007, appendix A). This gives

$$\mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}} L_j) = \frac{\alpha_j \mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}}) - D_{z_j}\mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}})}{(\alpha_0 + \alpha_1 + ... + \alpha_J)}$$

and I have found an expression for $\mathbf{L}^{\boldsymbol{\alpha}+\mathbf{e}_j}$. If $j = 0$ then use the fact that $\sum_k L_k = 1$ to write

$$\mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}} L_0) = \mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}}) - \sum_{j \neq 0} \mathbf{E}(\mathbf{L}^{\boldsymbol{\alpha}} L_j),$$

which is an expression for $\mathbf{L}^{\boldsymbol{\alpha}+\mathbf{e}_j}$. This completes the induction step, I have shown the collection $\{\overline{V}_j(\theta, G)(f) : f \in A(\mathcal{L})\}$ can be derived from $s_0(\mathbf{z}; \theta, G)$. Next I will show that $A(\mathcal{L})$ is determining.

Let $\mathbf{1}$ be a vector of ones and consider the collection $\mathbf{L}^{\mathbf{1}}\mathcal{L} = \{\mathbf{L}^{\mathbf{1}} L_j : j = 1, ..., J\}$, then $A(\mathbf{L}^{\mathbf{1}}\mathcal{L}) \subset A(\mathcal{L})$. It turns out that $A(\mathbf{L}^{\mathbf{1}}\mathcal{L})$ satisfies the conditions of the determining class theorem. This of course will imply that $A(\mathcal{L})$ is determining.

For the rest of this proof, without loss of generality, I set $\mathbf{z}_0 = 0$ and suppress it from the notation. Clearly $A(\mathbf{L}^{\mathbf{1}}\mathcal{L})$ is an algebra. To show that if $f \in A(\mathbf{L}^{\mathbf{1}}\mathcal{L})$ then f vanishes at infinity it is enough to consider finite products of functions in the set $\mathbf{L}^{\mathbf{1}}\mathcal{L}$. Notice that if f is one such function then $f \leq \mathbf{L}^{\mathbf{1}} L_j$ for some j. So it will be enough to show that $\mathbf{L}^{\mathbf{1}} L_j$ vanishes at infinity for every j. Fix j, I need to show that for any given ε there exists and M such that if $\|\overline{\mathbf{V}}\| \geq M$ then $\mathbf{L}^{\mathbf{1}} L_j(\overline{\mathbf{V}}) \leq \varepsilon$.

Now, for each $k \in J$ choose $M_{k,0}$ such that if $\overline{V}_k \geq M_{k,0}/J^{1/2}$ then $L_0(\overline{\mathbf{V}}) < \varepsilon$. Next, choose $M_{k,1}$ such that if $\overline{V}_k \leq -M_{k,1}/J^{1/2}$ then $L_k(\overline{\mathbf{V}}) \leq \varepsilon$. Now set $M = \max_{i,k} M_{k,i}$. Then this is the M we are looking for since if $\overline{\mathbf{V}} \geq M$ then $|\overline{V}_k| \geq M/J^{1/2}$ for some k. But this implies $\overline{V}_k \geq M/J^{1/2}$ or $\overline{V}_k \leq -M/J^{1/2}$. By choice of M this implies $\overline{V}_k \geq M_{k,0}/J^{1/2}$ or $\overline{V}_k \leq -M_{k,1}/J^{1/2}$. By choice of $M_{k,0}$ and $M_{k,1}$ this implies that $L_0(\overline{\mathbf{V}}) \leq \varepsilon$ or $L_k(\overline{\mathbf{V}}) \leq \varepsilon$. This implies $\mathbf{L}^{\mathbf{1}} L_j(\overline{\mathbf{V}}) \leq \varepsilon$.

Next I show that the collection $\mathbf{L}^{\mathbf{1}}\mathcal{L}$ separates points in R^J. This will be true if the collection $\mathcal{L}$ separates points in R^J, but this is obvious. Since the functions in $\mathbf{L}^{\mathbf{1}}\mathcal{L}$ are never zero, the third condition follows easily. This verifies the conditions of the determining class theorem for $A(\mathbf{L}^{\mathbf{1}}\mathcal{L})$. Therefore $A(\mathcal{L})$ is a determining class, the proof is complete. □

This concludes the section on identification. It will be more clear in the next section how these results are useful for constructing consistent estimators.

5. Consistency

Recall that the collection $\mathcal{C}$ is made up of indicator functions. In the typical example of a mixing model the collection $\mathcal{C}$ would be made up of smooth densities, in particular functions that vanish at infinity. Kiefer and Wolfowitz do not make this assumption explicitly, so one might hope that their methods could be applied directly in the multinomial choice setting. The following example shows that this is not a possibility.

5.1. **A counterexample.** The problem, as illustrated in the example, requires the dimensionality of randomness to be greater than 1. This is because, when $d_\zeta = 1$, we only have to add two points at infinity to keep the class $\mathcal{C}$ well behaved (this coincides with what Kiefer and Wolfowitz do). When $d_\zeta > 1$ many more points have to be added. Recall that the metric used in Kiefer and Wolfowitz is $d_{K-W}(G_1, G_2) = \int |G_1(x) - G_2(x)| \exp(-\|x\|)dx$.

Example 6. I use the structure from linear utility distribution free model (it would work equally well using the logit model). Set $J = 1$ and $d_\zeta = 2$. I will show that $\mathbf{s}(\mathbf{X}_t, Z_t; G)$ is not continuous in G with respect to d_{K-W} with positive probability according to $F_{\mathbf{x},z}$. To simplify the arguments, set $Z_t = 0$, the argument is analogous for all other values for Z_t. Then

$$\mathbf{s}_1(\mathbf{X}_t; G) = \int 1\{\mathbf{X}'_{1,t}\boldsymbol{\zeta} \geq 0\} dG(\boldsymbol{\zeta}).$$

Notice that if $\|\boldsymbol{\zeta}\|$ goes to infinity the kernel $1\{\mathbf{X}'_{1,t}\boldsymbol{\zeta} \geq 0\}$ converges to either 1 or 0 depending on the direction $\boldsymbol{\zeta}$ takes. Let l_1 and l_2 be two straight lines crossing the origin that are not parallel and that lie partially in the first quadrant. Let ζ^{n1} and ζ^{n2} be two sequences of points that converge to infinity along l_1 and l_2 respectively in the positive direction. Finally, let δ_{n1} (δ_{n2} resp.) be the sequence of dirac measures that put unit mass at ζ^{n1} (ζ^{n2} resp.) with distribution function G_{n1} (G_{n2} resp.).

Now, δ_{n1} and δ_{n2} converge to the same limit point in the completion of $\mathcal{P}(R^2)$ under the metric d_{K-W} above. To see this, notice that $G_{n1}(\boldsymbol{\zeta}) = G_{n2}(\boldsymbol{\zeta}) = 0$ unless $\|\boldsymbol{\zeta}\| \geq \|(\min\{\zeta_2^{n1}, \zeta_2^{n2}\}, \min\{\zeta_2^{n1}, \zeta_2^{n2}\})\| \to \infty$.

On the other hand, with positive probability $F_{\mathbf{x}}$ (recall $F_{\mathbf{x}}$ is absolutely continuous wrt lebesgue measure), $\mathbf{X}_{1,t} \in R^2$ is such that $|s_1(\mathbf{X}_{1,t}; \delta_{n1}) - s_1(\mathbf{X}_{1,t}; \delta_{n2})| \to 1$. Indeed, this holds true if $\mathbf{X}_{1,t} \in R^2$ is any vector with orthogonal subspace that lies between l_1 and l_2. ★

In this paper I solve these issues by completing the space $\mathcal{P}(R^{d_\zeta})$ in such a way that the market shares are continuous 'at infinity' for almost every value in $\text{supp}(\mathbf{X}_t, \mathbf{Z}_t)$.

5.2. **Some results.** The following is a general consistency theorem except that continuity and identification have been stated in terms of multinomial choice mixing problems. The corollaries that follow are about the linear utility specification.

Theorem 3. Consider the following assumptions

(1) the data $(\mathbf{X}_t, \mathbf{Z}_t)$ are i.i.d.;

(2) $\mathcal{P}$ is compact;

(3) $\Pr[\mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t; G) \neq \mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t; G_0)] > 0$ if $G \in \mathcal{P}$ and $G \neq G_0$ (identification);

(4a) $G\Big\{ \partial\{ \boldsymbol{\eta} \in H \mid C_j(\mathbf{X}_t, \mathbf{Z}_t; \boldsymbol{\eta}) = 1 \} \Big\} = 0$ for all j, for all $G \in \mathcal{P}$, for almost every $(\mathbf{X}_t, \mathbf{Z}_t)$ (continuity); or

(4b) $G\Big\{ \partial\{ \boldsymbol{\eta} \in H \mid C_0(\mathbf{X}_t, \mathbf{Z}_t; \boldsymbol{\eta}) = 1 \} \Big\} = 0$ for all $G \in \mathcal{P}$, for almost every $(\overline{\mathbf{X}}_t, \mathbf{Z}_t)$ where $\mathbf{X}_t = (1, \overline{\mathbf{X}}_t')$ (continuity).

If 1-3 and 4a hold, then for every $\varepsilon > 0$

$$\lim_{T \to \infty} \Pr\left\{ d_{\mathcal{P}}(\hat{G}_T^{FML}, G_0) \right\} = 0.$$

If 1-3 and 4b hold then for every $\varepsilon > 0$

$$\lim_{T \to \infty} \Pr\left\{ d_{\mathcal{P}}(\hat{G}_T^{PML}, G_0) \right\} = 0.$$

For a proof see appendix B.

This theorem is similar to the one that originally appeared in Kiefer and Wolfowitz (1956). Assumption 1 is used in their paper as well and, in both cases, can be relaxed. Assumption 2 is similar to the assumption that is used in their paper, the difference being that I do not restrict $\mathcal{P}$ to equal the completion of $\mathcal{P}(R^{d_\eta})$ under d_{K-W}. This will be important below. Condition 3 is analogous the their identification condition. Condition 4 corresponds to their continuity assumption. It differs from their continuity assumption in that I have tailored it to the multinomial choice framework. This modification was also used in Ichimura and Thompson (1999). Kiefer and Wolfowitz (1956) impose a 5th condition as well, that the criterion does not take the value $-\infty$ at the true parameter. In the multinomial choice setup this condition is trivially satisfied because $x\log(x)$ is bounded for $x \in [0,1]$.

The theorem leads to the following two corollaries about consistency of the npml estimator for a mixing distribution G in the linear utility models from examples 3 and 4.

Corollary 1. In the linear utility distribution free model assume (M1) and (M2) hold. Also assume: (i) conditional on $\mathbf{X}_t$, $\text{supp}(\mathbf{Z}_t) = R^J$; (ii) The marginal distribution of $\mathbf{X}_{j,t}$ contains a neighborhood of the origin for some j; (iii) The r. vector $(\mathbf{X}_t, \mathbf{Z}_t)$ is i.i.d.; (iv) The d.f. of $(\mathbf{X}_t, \mathbf{Z}_t)$, $\mathrm{F}_{\mathbf{xz}}$ has a density with respect to lebesgue measure. Then the npfml estimator for the mixing distribution $\widehat{\mathrm{G}}_t^{FML}$ converges weakly to G_0 in probability.

Corollary 2. In the linear utility distribution free model assume (M1) and (M2) hold. Also assume: (i) conditional on $\mathbf{X}_t$, $\text{supp}(\mathbf{Z}_t) = R^J$; (ii) $\mathbf{X}_{j,t} = (1, \overline{\mathbf{X}}_{j,t}\ ')'$ and $\overline{\mathbf{X}}_{j,t} \in R^{(d_\zeta - 1)}$ has full support in $R^{d_{(\zeta}-1)}$ for some j ; (iii) The r. vector $(\mathbf{X}_t, \mathbf{Z}_t)$ is i.i.d.; (iv) $\mathbf{X}_t = (\mathbf{1}, \overline{\mathbf{X}}\ ')')$ and the d.f. of the r. vector $(\overline{\mathbf{X}}, \mathbf{Z}_t)$ has a density with respect to lebesgue measure. Then the nppml estimator for the mixing distribution $\widehat{\mathrm{G}}_t^{PML}$ converges weakly to G_0 in probability.

Corollary 3. In the linear utility distribution free model assume (M1) and (M2) hold. Also assume: (i) Conditional on $\mathbf{X}_t$, supp($\mathbf{Z}_t$) contains an open set; (ii) The marginal distribution of $\mathbf{X}_{j,t}$ contains a neighborhood of the origin for some j ; (iii) $(\mathbf{X}_t, \mathbf{Z}_t)$ are i.i.d.; (iv) $\mathrm{F}_{\mathbf{xz}}$ has a density with respect to lebesgue measure. Then the npfml estimator for the mixing distribution $\widehat{\mathrm{G}}_t^{FML}$ converges weakly to G_0 in probability.

Corollary 4. In the linear utility distribution free model assume (M1) and (M2) hold. Also assume: (i) Conditional on $\mathbf{X}_t$, supp($\mathbf{Z}_t$) contains an open set; (ii) $\mathbf{X}_{j,t} = (1, \overline{\mathbf{X}}_{j,t}\ ')'$ and $\overline{\mathbf{X}}_{j,t} \in R^{(d_\zeta - 1)}$ has full support in $R^{d_(\zeta - 1)}$ for some j; (iii) $(\mathbf{X}_t, \mathbf{Z}_t)$ are i.i.d.; (iv) $\mathbf{X}_t = (\mathbf{1}, \overline{\mathbf{X}}\ ')')$ and the d.f. of the r. vector $(\overline{\mathbf{X}}, \mathbf{Z}_t)$ has a density with respect to lebesgue measure. Then the nppml estimator for the mixing distribution $\widehat{\mathrm{G}}_t^{PML}$ converges weakly to G_0 in probability.

Some remarks are in order. Because the identification results from the last section are joint across j I could have weakened condition 2 substantially in corollaries 1 and 3 at the cost of clarity. This could be an interesting direction to pursue for future research. Second, with an additional very weak assumption one could modify the theorem and the corollaries so that a sieve space could be used Θ_n. For example the sieve estimation strategy discussed by Bajari, Fox, Kim, and Ryan (2009b) could be rigorously justified using the framework from this paper. Below I will show how to prove the corollaries. Before I can do that I need to give a complete description of the models.

5.3. **Examples 3 and 4, continued.** Recall that in these examples I used cR^{d_ζ}, some compactification of R^{d_ζ}, to define H. The reason I work on this compactified space is so that condition 2 from the above theorem will be satisfied. So far, I have not assigned values to the function $\mathbf{C}$ when $\boldsymbol{\zeta}$ was a point at infinity (i.e., $\boldsymbol{\zeta} \in rR^{d_\zeta}$). This will be done once I have formally defined cR^{d_ζ}.

The trick will be to define cR^{d_ζ} and C on rR^{d_ζ} in such a way that condition 4 from the theorem still holds. Note that condition 4 is already true for any

distribution G that assigns probability 1 to the set R^{d_ζ}. So for $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in R^{d_\zeta}$ let

$$d_s(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = \arccos\left[\frac{(\boldsymbol{\zeta}_1', 1)(\boldsymbol{\zeta}_2', 1)'}{\|(\boldsymbol{\zeta}_1', 1)'\|\|(\boldsymbol{\zeta}_2', 1)'\|}\right]$$

define a metric on R^{d_ζ}. Then R^{d_ζ} is totally bounded. Let cR^{d_ζ} be the completion of R^{d_ζ} under this metric. Then cR^{d_ζ} is compact. This will be the compactification of R^{d_ζ} that will be used in both examples 3 and 4.

Example 3, continued. Recall that example 3 is the linear utility distribution free model. Now that I have described cR^{d_ζ} I can define $\mathbf{C}(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta})$ on the remainder rR^{d_ζ}. In completing the space I added one point at infinity for each $\boldsymbol{\zeta} \in R^{d_\zeta}$ with $\|\boldsymbol{\zeta}\| = 1$, so identify these points and let $\boldsymbol{\zeta}_0 \in R^{d_\zeta}$ with $\|\boldsymbol{\zeta}_0\| = 1$ be arbitrary. Then $\mathbf{C}(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}_0)$ is well defined. If $\boldsymbol{\zeta}_\infty$ is the point at infinity corresponding to $\boldsymbol{\zeta}_0$ then define

$$\mathbf{C}(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}_\infty) \equiv \mathbf{C}(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}_0).$$

★

In Example 4 the situation is the same.

Example 4, continued. Recall that example 4 is the linear utility logit model. With the notation from the previous example, define

$$\mathbf{C}(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}_\infty, \boldsymbol{\varepsilon}) \equiv \mathbf{C}(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}_0, \mathbf{0}).$$

★

5.4. **Some lemmas.** The proofs of the corollaries are long. It makes sense to break out some key statements in the form of lemmas. Hopefully this will make things easier to digest. The lemmas are proved in appendix C. As motivation for the lemmas that follow, notice that in the linear utility model the event $\partial\{\ \zeta \mid C_j(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) = 1\ \}$ (a tie) occurs only on sets of the form

$$\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid u_j(\mathbf{x}_j, z_j; \boldsymbol{\zeta}) = u_k(\mathbf{x}_k, z_k; \boldsymbol{\zeta})\}}$$

for some j, k. These are sets of the form

$$\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} + b = 0\}}.$$

From this, and in light of condition 4, one can see immediately that it would be useful to know just 'how many' hyperplanes can have positive probability for a fixed borel measure G.

Let $d_\zeta \geq 1$ be arbitrary. Consider the collection of $(d_\zeta - 1)$-dimensional hyperplanes in R^{d_ζ}. We can index these hyperplanes by $(\mathbf{a}, b) \in R^{(d_\zeta+1)}$. Let $H_{(\mathbf{a},b)} \equiv \{\mathbf{w} \in R^{d_\zeta} \mid \mathbf{a}'\mathbf{w} + b = 0\}$ and let G be an arbitrary finite measure on $(R^{d_\zeta}, \mathcal{B})$. Finally, let $\lambda^{(d_\zeta+1)}$ be lebesgue measure on $R^{d_\zeta+1}$. The following lemma says that 'most' hyperplanes in R^{d_ζ} have measure zero under G.

Lemma 1. Using the notation from above, for any fixed finite measure G on $(R^{d_\zeta}, \mathcal{B})$, we have $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid G(H_{(\mathbf{a},b)}) > 0\}) = 0$.

This lemma will be useful for ruling out boundaries in rR^{d_ζ} with positive probability under G for almost every $(\mathbf{X}_t, \mathbf{Z}_t)$. The next lemma allows me to rule out boundary points (e.g., ties) in rR^{d_ζ} with positive probability under G for almost every $(\mathbf{X}_t, \mathbf{Z}_t)$. Define $rH_{\mathbf{a}} = \overline{\{\zeta \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} + b = 0\}} \cap rR^{d_\zeta}$, where the closure operation is taken in the space cR^{d_ζ}. Note this set is independent of b. Finally, let λ^{d_ζ} be lebesgue measure on R^{d_ζ}. Then the following lemma is true.

Lemma 2. Using the notation as above, for any fixed finite measure G on $(cR^{d_\zeta}, \mathcal{B}_s)$, where now $\mathcal{B}_s$ is the borel sigma field on cR^r generated by d_s, we have $\lambda^{d_\zeta}(\{\mathbf{a} \in R^{d_\zeta} \mid G(rH_{\mathbf{a}}) > 0\}) = 0$.

In the case that $\mathbf{X}_t = (1, \mathbf{X}_{t,-1}\ ')'$ I will need a modification of lemmas 1 and 2. The proof will be analogous to the previous lemma, only I will need make an additional observation. In the following two lemmas $\lambda^{(d_\zeta-1)}$ is lebesgue measure on $R^{(d_\zeta-1)}$.

Lemma 3. Using the notation as above, for any fixed finite measure G on $(cR^{d_\zeta}, \mathcal{B}_s)$, where again $\mathcal{B}_s$ is the borel sigma field on cR^r generated by d_s, we have $\lambda^{d_\zeta}(\{(\mathbf{a}, b) \in R^{d_\zeta} \mid G(H_{(1,\mathbf{a}),b}) > 0\}) = 0$.

Lemma 4. Using the notation as above, for any fixed finite measure G on $(cR^{d_\zeta}, \mathcal{B}_s)$, where again $\mathcal{B}_s$ is the borel sigma field on cR^r generated by d_s, we have $\lambda^{(d_\zeta - 1)}(\{\mathbf{a} \in R^{(d_\zeta - 1)} \mid G(rH_{(1,\mathbf{a})}) > 0\}) = 0$.

5.5. **Proof of the corollary 1.** I now verify the conditions of Theorem 3 in the linear utility distribution free example. The other corollaries can be proved in a similar way. See appendix D.

proof of Corollary 1. Condition 1 is assumed directly. Let $d_{\mathcal{P}}$ metrize the weak-star topology for $\mathcal{P}(cR^{d_\zeta})$ relative to the space of d_s continuous bounded functions. Then $(\mathcal{P}(cR^{d_\zeta}), d_{\mathcal{P}})$ is compact (see Parthasarathy (1967), Theorem II.6.4). This establishes condition 2.

Given the support conditions on $\text{supp}(\mathbf{X}_t)$, and if $\mathbf{Z}_t$ has full support conditional on $\mathbf{X}_t$, then condition 3 holds on the set $\{\text{G} \in \mathcal{P}(cR^{d_\zeta}) \mid \text{G}(rR^{d_\zeta}) = 0\}$. To see this note that by the above identification result $\Pr[\mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t, \theta, G) \neq \mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t, \theta_0, G_0) \mid \mathbf{X}_t = \mathbf{x}] > 0$. In fact this is true with positive probability under $\mathbf{X}_t$ (this is enough to verify condition 3, see Folland (1999), Proposition 2.23). Say not, then by the assumptions on the support of $\mathbf{X}_t$ for $G \neq G_0$, we have $\mathbf{x}'\boldsymbol{\zeta} =^D \mathbf{x}'\boldsymbol{\zeta_0}$ for almost every $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$. The Cramér-Wold device implies $G = G_0$, a contradiction.

Now suppose G is such that $G(rR^{d_\zeta}) > 0$. Fix $\mathbf{x}$, then for some j

$$\int_{rR^{d_\zeta}} \mathbf{C}_j(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) dG(\boldsymbol{\zeta}) = \int_{rR^{d_\zeta}} \mathbf{C}_j(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}) dG(\boldsymbol{\zeta}) > \varepsilon$$

for all $\mathbf{z}$. Thus, for all $\mathbf{z}$, $s_j(\mathbf{x}, \mathbf{z}, G) > \varepsilon$. On the other hand, for any $\varepsilon > 0$ there exists a $\mathbf{z}$ far enough from the origin such that $s_j(\mathbf{x}, \mathbf{z}, G_0) < \varepsilon$. Therefore $\Pr[\mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t, \theta, G) \neq \mathbf{s}(\mathbf{X}_t, \mathbf{Z}_t, \theta_0, G_0) \mid \mathbf{X}_t = \mathbf{x}] > 0$ for every $\mathbf{x}$. Then because

the marginal distribution of $\mathbf{X}_t$ has a density with respect to lebesgue measure, condition 3 is satisfied (again, see Folland (1999), Proposition 2.23).

Condition 4a is the most involved. Its verification will illuminate the difficulty in generalizing the linear utility model to nonlinear utility examples. This is where I will need the lemmas from the last section.

Notice that the set $\partial\{\ \boldsymbol{\zeta} \mid C_j(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) = 1\ \}$ occurs only on sets of the form

$$B_1(\mathbf{a}, b) \equiv \overline{\{\zeta \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} + b = 0\}},$$

I will show that for every G fixed, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_1(\mathbf{a}, b)) = 0$. This follows from the lemmas.

According to lemma 1, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_1(\mathbf{a}, b) \cap R^{d_\zeta}) = 0$. Also, by lemma 2, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_1(\mathbf{a}, b) \cap rR^{d_\zeta}) = 0$. Putting these facts together we see that, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_1(\mathbf{a}, b)) = 0$. Because we assumed $(\mathbf{X}_t, \mathbf{Z}_t)$ has a density with respect to lebesgue measure, and because the $(\mathbf{a}, b)$ in the definition of the set $B_1(\mathbf{a}, b)$ is obtained by projection and convolution of $(\mathbf{X}_t, \mathbf{Z}_t)$, we can conclude that $G\Big\{\ \partial\{\boldsymbol{\zeta} \mid C_j(\mathbf{X}_t, \mathbf{Z}_t; \boldsymbol{\zeta}) = 1\}\Big\}$ almost everywhere $(\mathbf{X}_t, \mathbf{Z}_t)$. Condition 4a has been verified and the proof of the corollary is complete. □

6. Conclusion

In this paper I considered identification and consistent estimation in models of multinomial choice. I discussed a new and very general identification result that applies to multinomial choice models with an additive independent logistically distributed error term in the utility for choice j. I also described two new consistency results for nonparametric maximum likelihood estimation in the linear utility random coefficients multinomial choice models. In particular I showed how to fix the consistency theorem from Ichimura and Thompson (1999).

7. Appendix A: identification

proof of Theorem 2. By theorem 1 $\mathcal{A}$ is dense in $C_0(R^w)$ under sup norm. I first note that by the Riesz representation theorem for locally compact spaces $C_0(R^w)$ is determining for the class of all finite regular borel measures on R^w (see Conway (1990), page 75). In fact this implies that $\mathcal{A}$ is determining as well.

If $f \in C_0(R^w)$ is arbitrary and G is an arbitrary finite borel measure, then by the denseness of $\mathcal{A}$ there exists a sequence $f_n \in \mathcal{A}$ such that $\|f - f_n\|_\infty \to 0$. Since $|G(f) - G(f_n)| \leq \|f - f_n\|_\infty$ we can conclude $|G(f) - G(f_n)| \to 0$. Now, using the definition of a determining class we see that if $G(f)$ is known for every $f \in \mathcal{A}$ then by the statements made above $G(f)$ is known for every $f \in C_0(R^w$. Because $C_0(R^w$ is determining G is identified in the class of finite borel measures. The proof is complete. □

8. Appendix B: consistency

In this section I prove theorem 3. The proof is a modification of the proof that appears in Kiefer and Wolfowitz (1956). First some notation. Use $Y_{j,t}$ to denote the observed shares. Let $W_t = (\mathbf{Y}_t', \mathbf{X}_t', \mathbf{Z}_t')'$, $p(\mathbf{W}_t, G) = \prod_{j=0}^{J} s_j(\mathbf{X}_t, \mathbf{Z}_t, G)^{Y_{j,t}}$, and $q(\mathbf{W}_t, G) = s_0(\mathbf{X}_t, \mathbf{Z}_t, G)^{Y_{0,t}}(1-s_0(\mathbf{X}_t, \mathbf{Z}_t, G))^{(1-Y_{0,t})}$. Let $l(\mathbf{W}_t, G) = \log[p(\mathbf{W}_t, G)]$ and $m(\mathbf{W}_t, G) = \log[q(\mathbf{W}_t, G)]$. Also let $B(G, \rho) = \{G' \in \mathcal{P} \mid d_{\mathcal{P}}(G, G') \leq \rho\}$ and $p(\mathbf{W}_t, G, \rho) = \sup_{B(G,\rho)} p(\mathbf{W}_t, G)$. And let $q(\mathbf{W}_t, G, \rho) = \sup_{B(G,\rho)} q(\mathbf{W}_t, G)$. Then set $l(\mathbf{W}_t, G, \rho) = \log[p(\mathbf{W}_t, G, \rho)]$ and $m(\mathbf{W}_t, G, \rho) = \log[q(\mathbf{W}_t, G, \rho)]$. Let $\mathcal{P}(\delta) = \mathcal{P} \setminus B(G_0, \delta)$. Also, let $C_j(\mathbf{X}_t, \mathbf{Z}_t, G)$ denote the event that product j is chosen when the distribution is G.

The following lemmas establish that the appropriate objective functions are uniquely maximized.

Lemma 5. For each $G \in \mathcal{P}$, if $G \neq G_0$ then $\mathbf{E}[l(\mathbf{W}_t, G)] < \mathbf{E}[l(\mathbf{W}_t, G_0)]$.

Proof. It follows by Jensen's inequality that

$$\begin{aligned}\mathbf{E}\{\log[p(\mathbf{W}_t, G)/p(\mathbf{W}_t, G_0)]\} &\leq \mathbf{E}\{\log E[p(\mathbf{W}_t, G)/p(\mathbf{W}_t, G_0) \mid (\mathbf{X}_t, \mathbf{Z}_t)]\} \\ &= \mathbf{E}\{\log[1]\}\end{aligned}$$

note the inner expectation on the right side of the first line is with respect to G_0. By assumption 3 of Theorem 3 the inequality is strict. □

Lemma 6. For each $G \in \mathcal{P}$, if $G \neq G_0$ then $\mathbf{E}[m(\mathbf{W}_t, G)] < \mathbf{E}[m(\mathbf{W}_t, G_0)]$.

Proof. It follows by Jensen's inequality that

$$\begin{aligned}\mathbf{E}\{\log[q(\mathbf{W}_t, G)/q(\mathbf{W}_t, G_0)]\} &\leq \mathbf{E}\{\log E[q(\mathbf{W}_t, G)/q(\mathbf{W}_t, G_0) \mid (\mathbf{X}_t, \mathbf{Z}_t)]\} \\ &= \mathbf{E}\{\log[1]\}\end{aligned}$$

note the inner expectation on the right side of the first line is with respect to G_0. By assumption 3 of Theorem 3 the inequality is strict. □

The next lemmas establish continuity of the respective criterion functions.

Lemma 7. Under condition 4a, for each $G \in \mathcal{P}$, $\lim_{\rho \to 0} l(\mathbf{W}_t, G, \rho) = l(\mathbf{W}_t, G)$ almost surely. Also, $\lim_{\rho \to 0} \mathbf{E}[l(\mathbf{W}_t, G, \rho)] = \mathbf{E}[l(\mathbf{W}_t, G)]$.

Proof. Given assumption 4a from Theorem 3, $C_j(\mathbf{X}_t, \mathbf{Z}_t, G)$ is a continuity set for G a.s., condition on this event. Say the conclusion of the theorem is false. Then there exists a sequence $\{\rho_n\}$ and a sequence $\{G_n\}$ such that $d_{\mathcal{P}}(G_n, G) < \rho_n$ with $\rho_n \to 0$ such that $\liminf_{n \to \infty} l(\mathbf{W}_t, G_n) > l(\mathbf{W}_t, G)$. Since G_n converges weakly to G it follows by the portmanteau lemma that $s_j(\mathbf{X}_t, \mathbf{Z}_t, G_n) \to s_j(\mathbf{X}_t, \mathbf{Z}_t, G)$, and hence $l(\mathbf{X}_t, \mathbf{Z}_t, G_n) \to l(\mathbf{X}_t, \mathbf{Z}_t, G)$. This is a contradiction. This shows that for almost every $\mathbf{W}_t$, $\lim_{\rho \to 0} l(\mathbf{W}_t, G, \rho) = l(\mathbf{W}_t, G)$.

That $\lim_{\rho \to 0} \mathbf{E}[l(\mathbf{W}_t, G, \rho)] = \mathbf{E}[l(\mathbf{W}_t, G)]$ follows by the dominated convergence theorem and the fact that L^1 is complete. □

Lemma 8. Under condition 4b, for each $G \in \mathcal{P}$, $\lim_{\rho \to 0} m(\mathbf{W}_t, G, \rho) = m(\mathbf{W}_t, G)$ almost surely. Also, $\lim_{\rho \to 0} \mathbf{E}[m(\mathbf{W}_t, G, \rho)] = \mathbf{E}[m(\mathbf{W}_t, G)]$.

Proof. The proof is the same as in the lemma above except that now, given the criterion in this case, we only require C_0 to be a continuity set. □

proof of Theorem 3. Theorem 3 contains two results. The second result is proved in exactly the same way as the first, so I will only prove the first. I will show that for any $\delta > 0$, there exists an $\eta(\delta)$ with $0 < \eta(\delta) < 1$ such that

$$\mathbf{Pr}\left\{ \sup_{G \in \mathcal{P}(\delta)} \prod_{t=1}^{T} [p(\mathbf{W}_t, G)/p(\mathbf{W}_t, G_0)] > \eta(\delta)^T \right\} \to 0 \quad \text{as} \quad T \to \infty.$$

Theorem 3a clearly follows from this result. The proof is analogous to the proof of lemma A.4 from Ichimura and Thompson (1998). I repeat it here because there is an error in that proof and it is useful to have a correct version of that theorem. The previous two lemmas imply that for each $G \in \mathcal{P}(\delta)$ there exists a $\rho_G > 0$ such that $\mathbf{E}[l(\mathbf{W}_t, G, \rho_G)] < \mathbf{E}[l(\mathbf{W}_t, G_0)]$. Now $\mathcal{P}(\delta)$ is closed and hence compact. Therefore the cover $\{B(G, \rho_G)\}$ has a finite subcover. Index this finite subcover by j and so that $\mathcal{P}(\delta) \subset \bigcup_{n=1}^{N} B(G_n), \rho_n)$. Now

$$\begin{aligned}
&\Pr\left\{ \sup_{G \in \mathcal{P}(\delta)} \prod_{t=1}^{T} \Big[p(\mathbf{W}_t, G)/p(\mathbf{W}_t, G_0) \Big] > \eta(\delta)^T \right\} \\
&\leq \Pr\left\{ \sum_{n=1}^{N} \prod_{t=1}^{T} \Big[p(\mathbf{W}_t, G_n, \rho_n)/p(\mathbf{W}_t, G_0) \Big] > \eta(\delta)^T \right\} \\
&\leq \sum_{n=1}^{N} \Pr\left\{ \prod_{t=1}^{T} \Big[p(\mathbf{W}_t, G_n, \rho_n)/p(\mathbf{W}_t, G_0) \Big] > \eta(\delta)^T/N \right\} \\
&\leq \sum_{n=1}^{N} \Pr\left\{ \frac{1}{T} \sum_{t=1}^{T} \Big[l(\mathbf{W}_t, G_n, \rho_n) - l(\mathbf{W}_t, G_0) \Big] > \log \eta(\delta) - (\log N)/T \right\} \\
&\leq \sum_{n=1}^{N} \Pr\left\{ \frac{1}{T} \sum_{t=1}^{T} \max\left\{ l(\mathbf{W}_t, G_n, \rho_n) - l(\mathbf{W}_t, G_0); -C \right\} - \mathbf{E}_{n,0}(C) \right. \\
&\qquad\qquad \left. > -\mathbf{E}_{n,0}(C) + \log \eta(\delta) - (\log N)/T \right\} \\
&\quad + \sum_{n=1}^{N} \Pr\left\{ \frac{1}{T} \sum_{t=1}^{T} -C > \log \eta(\delta) - (\log N)/T \right\}.
\end{aligned}$$

Here $\mathbf{E}_{n,0}(C) = \mathbf{E}\big[\max\big\{l(\mathbf{W}_t, G_n, \rho_n) - l(\mathbf{W}_t, G_0); -C\big\}\big]$. Now choose C large enough so that $\mathbf{E}_{n,0}(C) < 0$ for $n = 1, ..., N$. This is possible by the first lemma from this appendix and by assumption 3. Then given $\varepsilon > 0$ pick $\eta(\delta)$ close enough to one so that $-\mathbf{E}_{n,0}(C) + \log \eta(\delta) - (\log N)/T > \varepsilon$ for $n = 1, ..., N$ and T large enough. Then, taking C larger if necessary the second term in the last line can be made to equal zero. Using assumption 1 from Theorem 3 and the law of large numbers completes the proof. □

9. Appendix C: proofs of lemmas

proof of Lemma 1. Fix G arbitrary, the proof is by induction. Notice that for any fixed $\mathbf{w} \in R^{d_\zeta}$ we have $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid \mathbf{w} \in H_{(\mathbf{a},b)}\}) = 0$. By a standard argument there are at most a countable number of points $(M_{0,i})_{i\in N}$, say, with positive probability under G. Therefore $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_{0,i} \in H_{(\mathbf{a},b)} \text{ for some } i\}) = 0$.

Next, consider all the lines (hyperplanes of dimension 1 in R^{d_ζ}). Notice that for any fixed line, say M_1, we have $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_1 \subset H_{(\mathbf{a},b)}\}) = 0$. Consider the collection of lines $\mathcal{R}_1 \equiv \{M_1 \mid \{(M_{0,i})_{i\in N}\} \cap M_1 = \emptyset\}$. If $M_{1,1}, M_{1,2} \in \mathcal{R}_1$ are two distinct lines then either $M_{1,1} \cap M_{1,2} = \emptyset$ or $M_{1,1} \cap M_{1,2}$ is a point. Therefore $G(M_{1,1} \cap M_{1,2}) = 0$ by definition of $\mathcal{R}_1$. This implies that there are at most countably many lines in $\mathcal{R}_1$, say $(M_{1,i})_{i\in N}$, with positive probability under G.

Also notice that for any fixed $H_{(\mathbf{a},b)}$ and line $M_1 \in \mathcal{R}_1$ then either $M_1 \subset H_{(\mathbf{a},b)}$ or, by definition of $\mathcal{R}_1$, $G(M_1 \cap H_{(\mathbf{a},b)}) = 0$ (since in this case $M_1 \cap H_{(\mathbf{a},b)}$ is either empty or is a point in R^r). This, along with the observation from the beginning of the last paragraph, implies $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid G(M_{1,i} \cap H_{(\mathbf{a},b)}) > 0 \text{ for some } i\}) = 0$.

Next, consider all hyperplanes of dimension 2 in R^{d_ζ}. Notice that for any fixed hyperplane of dimension 2, say M_2, we have $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_2 \subset H_{(\mathbf{a},b)}\}) = 0$. Now consider the collection

$$\mathcal{R}_2 \equiv \{M_2 \mid \{(M_{0,i})_{i\in N}\} \cap M_2 = \emptyset \text{ and for all } i \; M_{1,i} \cap M_2 \neq M_{1,i}\}.$$

If $M_{2,1}, M_{2,2} \in \mathcal{R}_2$ are two distinct hyperplanes of dimension 2 then either $M_{2,1} \cap M_{2,2} = \emptyset$ or $M_{2,1} \cap M_{2,2}$ is a hyperplane of dimension strictly less than 2. Therefore $G(M_{2,1} \cap M_{2,2}) = 0$ by definition of $\mathcal{R}_2$. This implies that there are at most countably many lines in $\mathcal{R}_2$, say $(M_{2,i})_{i \in N}$, with positive probability under G.

Also notice that for any fixed $H_{(\mathbf{a},b)}$ and hyperplane of dimension 2, say $M_2 \in \mathcal{R}_2$ then either $M_2 \subset H_{(\mathbf{a},b)}$ or, by definition of $\mathcal{R}_2$, $G(M_2 \cap H_{(\mathbf{a},b)}) = 0$. The first possibility is obvious. The fact that, if the first possibility is not true then the second is true, is less obvious. Notice that if $M_2 \subset H_{(\mathbf{a},b)}$ is not true then $M_2 \cap H_{(\mathbf{a},b)}$ is either empty or is a hyperplane of dimension 0 or 1 in R^r. If $M_2 \cap H_{(\mathbf{a},b)}$ is empty then the conclusion is obvious. If $M_2 \cap H_{(\mathbf{a},b)}$ is a point then the conclusion follows by definition of $\mathcal{R}_2$. Finally, if $M_2 \cap H_{(\mathbf{a},b)}$ is a hyperplane of dimension 1, then by definition of $\mathcal{R}_2$ the set $M_2 \cap H_{(\mathbf{a},b)} \in \mathcal{R}_1$ and is not equal to $M_{1,i}$ for any i (again by definition of $\mathcal{R}_2$). Therefore, by definition of $\mathcal{R}_1$ we must have $G(M_2 \cap H_{(\mathbf{a},b)}) = 0$. So the statement is true.

This last statement, along with the observation that $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_{2,i} \subset H_{(\mathbf{a},b)}\}) = 0$ for all i, imply $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid G(M_{2,i} \cap H_{(\mathbf{a},b)}) > 0 \text{ for some } i\}) = 0$.

Next, consider all hyperplanes of dimension 3 in R^{d_ζ}. Notice that for any fixed hyperplane of dimension 3, say M_3, we have $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_3 \subset H_{(\mathbf{a},b)}\}) = 0$. Now consider the collection

$$\mathcal{R}_3 \equiv \{M_3 \mid \{(M_{0,i})_{i \in N}\} \cap M_3 = \emptyset \text{ and for all } i\ M_{2,i} \cap M_3 \neq M_{2,i} \text{ and for all } i\ M_{1,i} \cap M_3 \neq M_{1,i}\}.$$

If $M_{3,1}, M_{3,2} \in \mathcal{R}_3$ are two distinct hyperplanes of dimension 3 then either $M_{3,1} \cap M_{3,2} = \emptyset$ or $M_{3,1} \cap M_{3,2}$ is a hyperplane of dimension strictly less than 3. If $M_{3,1} \cap M_{3,2}$ is of dimension 0 then $G(M_{3,1} \cap M_{3,2}) = 0$ by definition of $\mathcal{R}_3$. If $M_{3,1} \cap M_{3,2}$ is of dimension 1 then $M_{3,1} \cap M_{3,2} \in \mathcal{R}_1$ and does not contain any $M_{1,i}$. Therefore, by definition of $\mathcal{R}_1$ and $\mathcal{R}_3$, $G(M_{3,1} \cap M_{3,2}) = 0$. Finally, if $M_{3,1} \cap M_{3,2}$ is of dimension 2 then $M_{3,1} \cap M_{3,2} \in \mathcal{R}_2$ and does not contain any $M_{2,i}$. Therefore, by

definition of $\mathcal{R}_2$ and $\mathcal{R}_3$, $G(M_{3,1} \cap M_{3,2}) = 0$. This implies that there are at most countably many hyperplanes of dimension 3 in $\mathcal{R}_3$, say $(M_{3,i})_{i \in N}$, with positive probability under G.

Notice that for any fixed $H_{(\mathbf{a},b)}$ and hyperplane of dimension 3, say $M_3 \in \mathcal{R}_3$ then either $M_3 \subset H_{(\mathbf{a},b)}$ or, by definition of $\mathcal{R}_3$, $G(M_3 \cap H_{(\mathbf{a},b)}) = 0$. This last statement, along with the observation that $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid M_{3,i} \subset H_{(\mathbf{a},b)}\}) = 0$ for all i, imply $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid G(M_{3,i} \cap H_{(\mathbf{a},b)}) > 0 \text{ for some } i\}) = 0$.

The proof follows by continuing this process inductively until a collection $\mathcal{R}_{(d_\zeta-1)}$ with $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid H_{(\mathbf{a},b)} \notin \mathcal{R}_{(d_\zeta-1)}\}) = 0$ has been found such that there are at most countably many hyperplanes of dimension $(d_\zeta - 1)$ in $\mathcal{R}_{(d_\zeta-1)}$, say $(M_{(d_\zeta-1),i})_{i \in N}$, with with $G(M_{(d_\zeta-1),i}) > 0$. Then

$$\mathcal{R} = \{(\mathbf{a}, b) \mid H_{(\mathbf{a},b)} \in \mathcal{R}_{(d_\zeta-1)} \text{ and } H_{(\mathbf{a},b)} \neq M_{(d_\zeta-1),i} \text{ for all } i \in N\}$$

is a set such that $\lambda^{(d_\zeta+1)}(\{(\mathbf{a}, b) \mid (\mathbf{a}, b) \notin \mathcal{R}\}) = 0$ and such that $G(H_{(\mathbf{a},b)}) = 0$ for all $(\mathbf{a}, b) \in \mathcal{R}$. The proof is complete. □

proof of Lemma 2. The sets $rH_{\mathbf{a}}$ are indexed by $\mathbf{a}$. Because these sets may intersect, however, I can't argue directly that there are only a countable number of them with positive measure under G. Instead I will use an induction argument.

Recall from before that the collection of points attached at infinity rR^{d_ζ} is isomorphic to the unit hypersphere in R^{d_ζ}. In what follows, it may be easier to identify these points in your mind.

Use S_i to denote a subspace of R^{d_ζ} of dimension i. For each such subspace there is a unique corresponding manifold at infinity $\mathrm{M}_{d_\zeta-(i+1)}$. In fact M_i is the manifold of dimension i of the form $\overline{\mathrm{S}^{\perp}_{d_\zeta-(i+1)}} \bigcap r\mathbb{R}^{d_\zeta}$ for a given subspace $\mathrm{S}_{d_\zeta-(i+1)}$ with $i < d_\zeta - 1$, where the closure operator is taken with respect to the space (cR^{d_ζ,d_s}). $\mathrm{M}_{d_\zeta-(i+1)}$ is, in a way, dual to the subspace S_i. Notice that the M_i have the nice property that if $\mathrm{M}^1_i \neq \mathrm{M}^2_i$ then $\mathrm{M}^1_i \bigcap \mathrm{M}^2_i$ is a set of the form M_{i-1}.

(base step) A standard argument shows that there are at most a countable number of distinct manifolds of dimension 0 at infinity with positive probability under

G. Denote these by $\{\mathrm{M}_0^i\}$ and let $\{\mathrm{S}_{d_\zeta-1}^i\}$ denote their corresponding subspaces of dimension $d_\zeta - 1$.

Now let $\mathcal{S}_{d_\zeta-2}$ be the collection of subspaces of dimension $d_\zeta - 2$ with $\mathrm{S}_{d_\zeta-2} \cap \{\mathrm{S}_0^i\} = \{\mathbf{0}\}$. If M_1^1 and M_1^2 are manifolds at infinity corresponding to distinct $\mathrm{S}_{d_\zeta-2}^1, \mathrm{S}_{d_\zeta-2}^2 \in \mathcal{S}_{d_\zeta-2}$ then $\mathrm{M}_1^1 \cap \mathrm{M}_1^2 = \mathrm{M}_0$ for some M_0. Therefore, by definition of $\mathcal{S}_{d_\zeta-2}$, we have $G(\mathrm{M}_1^1 \cap \mathrm{M}_1^2) = 0$.

Conclude that there are at most a countable number of $\mathrm{S}_{d_\zeta} \in \mathcal{S}_{d_\zeta-2}$ such that $G(\mathrm{M}_1) > 0$. Use $\{\mathrm{S}_{d_\zeta-2}^i\}$ to denote this collection. Continue inductively in this manner until we have found at most a countable number of subspaces of dimension 1 with $\mathrm{S}_1 \cap \{\bigcup_{k=2}^{d_\zeta-1} \bigcup_{i=1}^{\infty} \mathrm{S}_k^i\} = \{0\}$ that have corresponding manifolds at infinity with positive probability under G. Denote this collection by $\{\mathrm{S}_1^i\}$. Finally, notice that $\lambda^{d_\zeta}(\{\mathbf{a} \in R^{d_\zeta} \mid \mathbf{a} \in \bigcup_{k=1}^{d_\zeta-1} \bigcup_{i=1}^{\infty} \mathrm{S}_k^i\}) = 0$. Also, if $\mathbf{a} \notin \bigcup_{k=1}^{d_\zeta-1} \bigcup_{i=1}^{\infty} \mathrm{S}_k^i$, then $\mathrm{G}(rH_{\mathbf{a}}) = 0$. □

proof of Lemma 3. Lemma 3 follows directly from the proof of lemma 1. □

proof of Lemma 4. I will again use the same induction argument that I used to prove lemma 2 to obtain $\{\mathrm{S}_k^i\}$ with $i \in N$ and $k \in \{1, ..., d_\zeta - 1\}$. This time we notice that $\lambda^{(d_\zeta-1)}(\{\mathbf{a} \in R^{(d_\zeta-1)} \mid (1, \mathbf{a}')' \in \bigcup_{k=1}^{d_{(\zeta}-1)} \bigcup_{i=1}^{\infty} \mathrm{S}_k^i\}) = 0$. Again, by construction, if $(1, \mathbf{a}')' \notin \bigcup_{k=1}^{d_\zeta-1} \bigcup_{i=1}^{\infty} \mathrm{S}_k^i$, then $\mathrm{G}(rH_{(1,\mathbf{a}')'}) = 0$. □

10. Appendix D: proofs of corollaries

proof of Corollary 2. Note that identification only required observations on $Y_{0,t}$. Therefore conditions 1-3 follow for the reasons explained above. Condition 4b is verified in a similar way, using Lemmas 3 and 4 instead of Lemmas 1 and 2. Notice that the set $\partial\{\ \boldsymbol{\zeta} \mid C_0(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) = 1\ \}$ occurs only on sets of the form

$$B_2(\mathbf{a}, b) \equiv \overline{\{\zeta \in R^{d_\zeta} \mid (1, \mathbf{a}')\boldsymbol{\zeta} + b = 0\}},$$

I will show that for every G fixed, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_1(\mathbf{a}, b)) = 0$. This follows from the lemmas.

According to lemma 3, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_2(\mathbf{a}, b) \cap R^{d_\zeta}) = 0$. Also, by lemma 4, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_2(\mathbf{a}, b) \cap rR^{d_\zeta}) = 0$. Putting these facts together we see that, for lebesgue almost every $(\mathbf{a}, b)$, $G(B_2(\mathbf{a}, b)) = 0$. Because we assumed $(\overline{\mathbf{X}}_t, \mathbf{Z}_t)$ has a density with respect to lebesgue measure, and because the $(\mathbf{a}, b)$ in the definition of the set $B_2(\mathbf{a}, b)$ is obtained by projection and convolution of $(\overline{\mathbf{X}}_t, \mathbf{Z}_t)$, we can conclude that $G\Big\{ \partial\{\boldsymbol{\zeta} \mid C_0(\mathbf{X}_t, \mathbf{Z}_t; \boldsymbol{\zeta}) = 1\}\Big\}$ almost everywhere $(\overline{\mathbf{X}}_t, \mathbf{Z}_t)$. Condition 4b has been verified and the proof of the corollary is complete. □

Verification of the conditions in the linear utility logit model is similar to the steps laid out above. A striking difference is that in this model we only need $\mathbf{Z}_t$ to have support containing some open set conditional on $\mathbf{X}_t$. The support conditions on $\mathbf{X}_t$ are the same as before.

proof of Corollary 3. Condition 1 is assumed. Again, let $d_{\mathcal{P}}$ metrize the weak-star topology for $\mathcal{P}(cR^{d_\zeta})$ relative to the space of d_s continuous bounded functions. Then $(\mathcal{P}(cR^{d_\zeta}), d_{\mathcal{P}})$ is compact. This establishes condition 2.

It is possible to verify the identification condition (condition 3) in this model because

$$\mathrm{s}_0(\mathbf{x}, \mathbf{z}; \mathrm{G}) \equiv \int_{R^{d_\zeta} \times R^J} \mathrm{C}_0(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}, \boldsymbol{\varepsilon}) d(G \times F_\varepsilon)(\boldsymbol{\zeta}, \boldsymbol{\varepsilon}) + \mathrm{K}(\mathbf{x}; G),$$

where $\mathrm{K}(\mathbf{x}; G)$ is independent of $\mathbf{z}$ and equals

$$\begin{aligned} \mathrm{K}(\mathbf{x}; G) &= \int_{rR^{d_\zeta} \times R^J} \mathrm{C}_0(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}, \mathbf{0}) d(G \times F_\varepsilon)(\boldsymbol{\zeta}, \boldsymbol{\varepsilon}) \\ &= \int_{rR^{d_\zeta}} \mathrm{C}_0(\mathbf{x}, \mathbf{0}; \boldsymbol{\zeta}, \mathbf{0}) d(G)(\boldsymbol{\zeta}). \end{aligned}$$

So that

$$\partial_z^{\mathbf{1}} s_0(\mathbf{x}, \mathbf{z}; \mathrm{G}) = C \int_{R^{d_\zeta}} \mathbf{L}^{\mathbf{1}}(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) dG(\boldsymbol{\zeta}),$$

where C is some constant. This demonstrates that the identification result for the logit model, proposition 2, only required observations on the outside option (the

distribution of $Y_{0,t}$). Note that if $\mathrm{K}(\mathbf{x}; G) \neq 0$ then $G(R^{d_\zeta}) < 1$. Condition 3 is confirmed using this observation, the continuity of s, and proposition 2 from the section on identification. Recall, proposition 2 allowed us to identify $\overline{\mathbf{V}}(\mathbf{x}; \boldsymbol{\zeta})$ within the collection of finite borel measures (not just probability measures).

Condition 4a is verified in a similar way. Again, boundary sets $\partial\{\ \boldsymbol{\zeta} \mid C_j(\mathbf{x}, \mathbf{z}; \boldsymbol{\zeta}) = 1\ \}$ can be divided into four types:

$$A_1(\mathbf{a}, b) \equiv \{(\boldsymbol{\zeta}, \boldsymbol{\varepsilon}) \in R^{d_\zeta} \times R^J \mid \mathbf{a}'\boldsymbol{\zeta} + b + \varepsilon_i - \varepsilon_j = 0\},$$
$$A_2(\mathbf{a}, b) \equiv \overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta} \times R^J$$

In fact, if $G \in \mathcal{P}$ then $G = G_1 \times F_\varepsilon$ where F_ε is a logistic probability measure. In particular F_ε is continuous. Therefore, conditional on $\boldsymbol{\zeta}$, the sets $A_1(\mathbf{a}, b)$ have zero probability $F_\varepsilon(A_1(\mathbf{a}, b)) = 0$. This implies the probability is zero for the unconditional distribution.

Next, note that $(G_1 \times F_\varepsilon)(A_2(\mathbf{a}, b)) = G_1(\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta})$. But $G_1(\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid \mathbf{a}'\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta}) = 0$ for lebesgue almost every $(\mathbf{a}, b)$ by lemma 2. Therefore condition 4.a holds and the proof is complete. □

proof of Corollary 4. Conditions 1-3 follow by the arguments made above. Condition 4b follows similarly because if $\mathbf{X}_t = (1, \overline{\mathbf{X}}_t\ ')'$ boundary sets are of the form

$$B_1(\mathbf{a}, b) \equiv \{(\boldsymbol{\zeta}, \boldsymbol{\varepsilon}) \in R^{d_\zeta} \times R^J \mid (1, \mathbf{a}')\boldsymbol{\zeta} + b + \varepsilon_i = 0\}$$
$$B_2(\mathbf{a}, b) \equiv \overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid (1, \mathbf{a}')\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta} \times R^J.$$

Again, if $G \in \mathcal{P}$ then $G = G_1 \times F_1$ where F_1 is a logistic probability measure. In particular F_1 is continuous. Therefore, conditional on $\boldsymbol{\zeta}$, the sets $B_1(\mathbf{a}, b)$ have zero probability: $F_\varepsilon(B_1(\mathbf{a}, b)) = 0$. This implies the probability is zero for the unconditional distribution.

Next, note that $(G_1 \times F_1)(B_2(\mathbf{a}, b)) = G_1(\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid (1, \mathbf{a}')\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta})$. But $G_1(\overline{\{\boldsymbol{\zeta} \in R^{d_\zeta} \mid (1, \mathbf{a}')\boldsymbol{\zeta} = 0\}} \cap rR^{d_\zeta}) = 0$ for lebesgue almost every $(\mathbf{a}, b)$ by lemma 4. Therefore condition 4.a holds and the proof is complete.

□

11. BIBLIOGRAPHY

Bajari, Fox, Kim, and Ryan (2007b). "A simple nonparametric estimator for the distribution of random coefficients in discrete choice models," Working paper.

Berry and Haile (2008). "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Working paper.

Billingsley, Patrick (1968). *Convergence of probability measures*, John Wiley & Sons, Inc., New York.

Chen, Xiaohong (2007). "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, (J.J. Heckman and E.E. Leamer, ed.), **6B**, Elsevier, North-Holland, Amsterdam. pp. 5549-5623.

Conway (1990). *A course in functional analysis*, Springer-Verlag, New York.

Dudley, R.M. (1999). *Uniform central limit theorems*, Cambridge University Press, New York.

Engelking, Ryszard (1989). *General topology*, Heldermann Verlang, Berlin.

Fedorchuk, V.V. (1998). "Functors of probability measures in topological categories," *Journal of Mathematical Sciences* **91**, pp. 3157-3204.

Folland (1999). *Real analysis: modern techniques and their applications*, John Wiley & Sons, Inc., New York.

Halmos, P.R. (1950). *Measure theory*, Van Nostrand, Princeton, NJ.

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

Kiefer and Wolfowitz (1956). "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics* **27** pp. 887-906.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces*, Academic Press, New York.

Ratcliffe J. G. (2006). *Foundations of hyperbolic manifolds*, Springer, New York.

Traine, K. (2009). *Discrete choice methods with simulation*, Cambridge University Press, New York.

Stroock, Daniel W. (1993). *Probability Theory: An analytic view*, Cambridge University Press, Cambridge.

van der Vaart and Wellner (2000). *Weak convergence and empirical processes: with applications to statistics*, Springer-Verlag, New York.

Wald, Abraham (1949). "Note on the consistency of the maximum likelihood estimate," *The Annals of Mathematical Statistics* **20**, pp. 595-601.

White, H., Wooldridge, J. (1991). "Some results on sieve estimation with dependent obserations," in *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*, (Barnett, W.A., Powell, J., Tauchen, G. ed.), Cambridge University Press, Cambridge. pp. 459-493.

Current address: New Haven, CT

E-mail address: eric.becker.m@gmail.com

URL: pantheon.yale.edu/~emb72