Moment Inequalities for Multinomial Choice with Fixed Effects

Ariel Pakes
Harvard University and NBER

Jack Porter
University of Wisconsin

August 15, 2014

Abstract

We propose a new approach to the semiparametric analysis of multinomial choice models with fixed effects and a group (or panel) structure. A traditional random utility framework is employed, and the key assumption is a group homogeneity condition on the disturbances. This assumption places no restrictions on either the joint distribution of the disturbances across choices or their within group (or across time) correlations. This work follows a substantial nonlinear panel literature (Manski 1987, Honore 1992, Abrevaya 1999, 2000) with the distinction that multiple covariate index functions now determine the outcome. A novel within-group comparison leads to a set of conditional moment inequalities that provide partial identifying information about the parameters of the observed covariate index functions, while avoiding the incidental parameter problem. We extend our framework to allow for: certain types of endogenous regressors (including lagged dependent variables and conditional heteroskedasticity), set-valued covariates, and parametric distributional information on disturbances.

1 Introduction

This paper proposes a new approach to semiparametric identification of multinomial choice models with fixed effects and a group (or panel) structure. We employ a traditional random utility framework (McFadden 1974) where utility is additively separable between unobservables, which include a disturbance and choice-specific fixed effects, and an index function of covariates and parameters. The key assumption is a group homogeneity condition on the disturbances. We place no restrictions on the joint distribution of the disturbances across choices. Under this specification, a novel within-group comparison leads to a set of conditional moment inequalities that provide identifying information about the parameters of the index functions. We then extend our framework to allow for set-valued covariates, certain types of endogenous regressors, and the incorporation of parametric distributional information on disturbances.

This work continues a substantial literature that has focused on extending nonlinear econometric models to allow for fixed effects while relaxing parametric distributional assumptions on disturbances. Manski (1987) applied his maximum score approach to the binary choice model with fixed effects. Honore (1992) further developed Powell's (1986) trimmed least squares approach to estimate the censored regression model with fixed effects. Abrevaya (1999) developed a new approach to estimation to allow for fixed effects in the transformation model, and further extended Han's (1987) generalized regression model to include fixed effects in Abrevaya (2000).

The multinomial choice setup considered in the current work presents an additional complexity relative to the models in this previous literature. In particular, the multinomial choice model depends on *multiple* index functions of the covariates, where each index function corresponds to a choice-specific random utility. The main insight of our identification strategy is that a comparison of the multiple index functions for any two within-group observations has observable implications on the relative likelihood of certain choice outcomes. To motivate this approach and connect to the previous literature, we begin by applying this idea to a general model with a *single* index function of the covariates, which we call the weakly monotone transformation model. Under the group homogeneity assumption on disturbances, the difference in outcome distributions from two within-group observations must come from differences in the single index functions for each observation. Weak monotonicity then implies that the difference in outcome distributions obeys a stochastic dominance relation which generates a set of moment inequalities. This derivation builds on the identification argument for binary choice in Manski (1987).

Having established this stochastic dominance result in the weakly monotone transformation model, we move to our main focus, the discrete choice random utility model with choice-specific fixed effects. This random utility model has appeared in the econometrics literature in many different guises. With panel data problems in mind, Chamberlain (1980) uses an assumption of logistic disturbances to provide a novel conditional likelihood method of identification and estimation. An alternative application is in the demand literature where markets are the grouping device, the within group observations are consumers, and the choice-specific fixed effects represent product level unobservables (e.g. Berry, Levinsohn, and Pakes 2004). Markets are also used as a grouping device when analyzing firm decision making (e.g. entry decisions) with the market-specific fixed effect representing unobserved determinants of the market's profitability (e.g. Pakes 2014).

The semiparametric version of the model considered here does not place parametric restrictions on the disturbance distribution. The only restriction on the disturbances is a group homogeneity assumption. To understand how this assumption is used consider two observations in the same group and the differences of their observable index functions for the different possible choices. Rank these differences from largest to smallest. We derive inequalities showing that the choices corresponding to the largest index function differences are more likely to be observed for the observation with the larger difference in indices. Such inequalities apply to each subset of the highest ranked differences in choices. The derived inequalities can be simply expressed as a set of conditional moment conditions satisfied by the true value of the index function parameter. These within-group moment inequalities can be seen as the analog to the within-group identification of continuous choice models and single-index weakly monotone transformation models. Finally, the derived conditional moment inequalities can be used for estimation and inference using methods developed in a recent literature (e.g. Andrews and Shi 2013, Armstrong 2011, Chetverikov 2011, Chernozhukov, Lee, and Rosen 2013, and Aradillas-López, Gandhi, and Quint 2013).

The main cost of the semiparametric flexibility in our model is that the conditional moment inequalities will, in general, only partially identify the index function parameters. On the other hand, there are several potential benefits to the semiparametric approach. First, the index function differencing step circumvents the need to account for the fixed effects (which may be correlated with the observed covariates). Hence this approach does not suffer from an incidental parameter problem when group sizes are small and the number of groups is large, as is often the case in panel data models.

Second, when working with multinomial choice applications we need not place any restrictions on the joint distribution of disturbances *across choices*. In particular, the marginal distributions of the choice-specific disturbances can differ arbitrarily across choices and can have bounded support or mass points. Perhaps more important, no restrictions are placed on either: (i) the covariance matrix of disturbances across choices (so the practitioner need not worry about vestiges of independence of irrelevant alternatives or limits on cross price elasticities), or (ii) their within-group (in panel data, across time) correlations. Though the group structure is required to deal with the fixed effects, the groups can be as small as two observations and the disturbance distribution can vary arbitrarily across groups.

Third, we show that straightforward modifications of our assumptions enable us to analyze models with certain forms of dependence of the disturbance distribution on the regressors; i.e. discrete choice models with endogenous regressors. These extensions include the analysis of models in which a lagged value of the dependent variable is a regressor. Since choice-specific fixed effects are included, this extension of our model allows for both "state dependence" *and* state-specific "heterogeneity". Additionally, we modify our group homogeneity assumption to show how control variables can be incorporated to handle certain kinds of conditional heteroskedasticity and other forms of regressor endogeneity.

Fourth, we explicitly allow for the situation where two or more choices have either the same random utility with positive probability or the same value for the index functions. We show in our extensions that this enables us to analyze discrete choice problems with set-valued regressors (which includes estimation problems with generated regressors).

Finally, the semiparametric estimator can be computed without calculating orthant probabilities. This lessens the computational burden of obtaining estimates, especially in choice models with rich disturbance distributions.

After analyzing the semiparametric problem we move to the case where the distribution of disturbances is known up to a finite dimensional parameter vector. No other restrictions are imposed on this distribution. In particular the group homogeneity assumption is not needed for the inequalities we derive for the parametric case. If the group homogeneity assumption is relevant the comparison between the parametric and semiparametric estimator should provide an indication of whether the parametric assumption is appropriate for the problem at hand.

Multinomial discrete choice estimation is extensively used in almost all fields that empirically analyze the determinants of agents' choices. Applications have typically employed parametric forms of the multinomial model. Manski (1975) introduced a semiparametric, maximum score approach to *point* identification and estimation for multinomial choice without choice-specific fixed effects. Assuming independent and identical distributions of the unobservable components of the different choices, Manski uses differences in the observable, parametric component of random utility across choices for identification. Using Manski's identification approach, Fox (2007) shows that exchangeability of the unobservable component across choices is sufficient for identification, and Yan (2013) obtains the limiting distribution for a smoothed version of the multinomial maximum score estimator. Lee (1995) provides an alternative semiparametric approach to multinomial choice for models without choice-specific fixed effects using an assumption of an i.i.d. distribution of disturbances across agents. Rather than imposing conditions on the joint distribution of the disturbances across choices our approach requires that the joint distribution of the choice-specific unobservables does not differ across observations in a group, but leaves the distribution of disturbances across choices unrestricted. The different assumptions are likely to be useful in different applications.

The paper is structured as follows. Section 2 illustrates the application of our approach to a large class of single index models. Section 3 begins with notation for the standard random utility model for multinomial choice with fixed effects. We then present and discuss our assumptions on the disturbance vector, and conclude with a derivation of the moment inequalities implied by those assumptions. In section 4, extensions of the basic framework to set-valued and endogenous regressors are considered. Section 5 analyzes the case where the disturbance distribution is indexed by a finite-dimensional parameter vector. The conclusion summarizes and notes implications for identification and estimation of other features of the model, such as marginal effects.

2 Weakly Monotone Transformation Model

To fix ideas, we begin by introducing the basic intuition behind our identification strategy in a weakly monotone transformation model. The model is a more general version of the generalized regression model introduced in Han (1987) and Abrevaya (2000). As we note below, there are many classic papers focusing on identification and semiparametric estimation of a variety of special cases of the model we describe. We are able to consider a quite general version of this model because we are not concerned with achieving point identification. Our interest in the following sections is a multinomial choice model, where the point identification arguments from special cases of the weakly monotone transformation model do not carry over. The advantage of starting with the weakly monotone transformation model is that it provides a familiar setting where the principal idea behind our partial identification approach can be made transparent without the complication of multiple index functions present in the multinomial choice setup.

We assume that the data has a group structure and let $i = 1, \ldots, n$ index different "groups" of observations, and $t = 1, \ldots, T_i$ index the different observations within a group i. This includes the traditional panel data situation where where (i, t) indexes individuals and time respectively.

The weakly monotone transformation model is:

$$y_{i,t} = \phi\Big(g(x_{i,t}, \beta_0), \lambda_i, \varepsilon_{i,t}\Big) \tag{1}$$

where $y_{i,t}$ is an observed scalar outcome variable, $x_{i,t}$ is an observed vector of covariates, λ_i is the (unobserved) fixed effect for group i, and $\varepsilon_{i,t}$ is an unobserved disturbance. The defining characteristic of this model is that ϕ is assumed to be weakly monotone increasing in its first argument, $g(x_{i,t}, \beta_0)$. We assume g is a known function that specifies a parametric, scalar index in the observed covariates. Typically, g will simply be chosen to have a linear form, $g(x_{i,t}, \beta_0) = x_{i,t}'\beta_0$. Interest centers on the parameter, β_0, of this index function. The function $\phi(\cdot)$ can be unknown, and the dimension of unobserved variables $(\lambda_i, \varepsilon_{i,t})$ is unrestricted.

Examples

Generalized Regression with Fixed Effects. A close relative of the model (1) is Abrevaya's (2000) fixed effects version of Han's (1987) generalized regression model. Abrevaya's model is: $y_{i,t} = D \circ F(x_{i,t}'\beta_0, \lambda_i, \varepsilon_{i,t})$, where D is weakly increasing and known and F is strictly increasing in its first and last arguments. The disturbance $\varepsilon_{i,t}$ is scalar. The generalized regression model can be considered a special case of the weakly monotone transformation model given in (1).

Binary Choice with Fixed Effects. Manski (1987) introduced maximum score estimation for the model $y_{i,t} = \mathbf{1}\{x_{i,t}'\beta_0 + \lambda_i + \varepsilon_{i,t} \geq 0\}$, where $\mathbf{1}\{A\}$ is an indicator function for event A.

Censored Regression with Fixed Effects. Honore (1992) extended Powell's (1986) censored regression model to include fixed effects: $y_{i,t} = \max\{0, x_{i,t}'\beta_0 + \lambda_i + \varepsilon_{i,t}\}$.

The only stochastic restriction we will consider is the following group homogeneity assumption.

Assumption M *Given the conditioning set $(x_{i,s}, x_{i,t}, \lambda_i)$, for any $s \neq t$, the conditional distributions of $\varepsilon_{i,s}$ and $\varepsilon_{i,t}$ are the same:*

$$\varepsilon_{i,s}\big|x_{i,s}, x_{i,t}, \lambda_i \ \sim\ \varepsilon_{i,t}\big|x_{i,s}, x_{i,t}, \lambda_i.$$

Assumption M places no restrictions on the joint distribution of $x_{i,s}$ (or $x_{i,t}$) and λ_i, so that arbitrary correlation between the fixed effects and the covariates is allowed. An equivalent restatement of Assumption M is $(\lambda_i, \varepsilon_{i,s})\big|x_{i,s}, x_{i,t} \ \sim\ (\lambda_i, \varepsilon_{i,t})\big|x_{i,s}, x_{i,t}$. From this equivalence, it is evident that the fixed effect could simply be absorbed into the notation of the disturbance, e.g. $u_{i,t} = (\lambda_i, \varepsilon_{i,t})$.[1] However, we will maintain the current notation to be clear about the inclusion of a fixed effect.

Manski (1987) includes an identical assumption to Assumption M. That paper, and others cited above, go on to provide sufficient conditions for identifying the parameter β_0 in different special cases of the model in equation (1). We suffice with the implications of Assumption M, weak monotonicity of ϕ in its first argument, and the assumption that g is known up to the parameter, β_0. In particular since $y_{i,t} = \phi(g(x_{i,t}, \beta_0), \lambda_i, \varepsilon_{i,t})$ and similarly for $y_{i,s}$, Assumption M implies that the only difference between the conditional distributions of $y_{i,t}$ and $y_{i,s}$ must come from differences in $g(x_{i,t}, \beta_0)$ and $g(x_{i,s}, \beta_0)$. The weak monotonicity of ϕ, then yields a stochastic dominance result. We formalize this argument as follows. Fix i and suppose $g(x_{i,s}, \beta_0) \geq g(x_{i,t}, \beta_0)$. Then for $y \in \mathbb{R}$,

$$\begin{aligned}
&\Pr(y_{i,s} \geq y \,|\, x_{i,s}, x_{i,t}, \lambda_i) = \Pr(\phi(g(x_{i,s}, \beta_0), \lambda_i, \varepsilon_{i,s}) \geq y \,|\, x_{i,s}, x_{i,t}, \lambda_i) \\
&\quad \geq \Pr(\phi(g(x_{i,t}, \beta_0), \lambda_i, \varepsilon_{i,s}) \geq y \,|\, x_{i,s}, x_{i,t}, \lambda_i) \qquad \text{since } g(x_{i,s}, \beta_0) \geq g(x_{i,t}, \beta_0) \\
&\quad = \Pr(\phi(g(x_{i,t}, \beta_0), \lambda_i, \varepsilon_{i,t}) \geq y \,|\, x_{i,s}, x_{i,t}, \lambda_i) \qquad \text{by Assumption } M \\
&\quad = \Pr(y_{i,t} \geq y \,|\, x_{i,s}, x_{i,t}, \lambda_i).
\end{aligned}$$

This gives us our first proposition.

Proposition 1 *Under Assumption M, two observations from the same group i with $g(x_{i,s}, \beta_0) \geq g(x_{i,t}, \beta_0)$ generated by the weakly monotone transformation model (1) must satisfy the first-order stochastic dominance[2] relation*

$$\Pr(y_{i,s} \geq y|x_{i,s}, x_{i,t}, \lambda_i) \geq \Pr(y_{i,t} \geq y|x_{i,s}, x_{i,t}, \lambda_i)$$

for all $y \in \mathbb{R}$.

One could turn this result into a conditional moment inequality that would define an identified set containing β_0. Our focus, however, is on the fact that this stochastic dominance

[1] Indeed in the monotone transform model all we require is that the function $\phi(\cdot)$ be weakly separable in (i) the index of covariates and parameters and (ii) a function of $(\lambda_i, \epsilon_{i,t})$. We could also allow both ϕ and g to vary with i. The transformation ϕ could differ with i in unknown ways as long as weak monotonicity is maintained. The index function g could also take on a different form for each group i, but each group form would need to be known.

[2] First-order stochastic dominance can be defined to include the condition that the probability inequality is strict for some value of y. In this sense, the conclusion of the proposition is a *weak* first-order stochastic dominance condition.

relationship is based only on the group homogeneity assumption and weak monotonicity. In particular no differencing was needed and hence this within-group variation result was achieved without the usual linearity assumptions. That is, the index function g is allowed to be nonlinear in form and the fixed effect is not required to be additive.

The next section extends the use of the group homogeneity assumption to multinomial choice to obtain a set of stochastic dominance conditions. In multinomial choice, the outcome is a function of separate covariate index functions, fixed effects, and disturbance terms for each choice. Group homogeneity will, again, ensure that the differences in the conditional distributions of two outcomes from the same group will be determined by differences in the covariate index functions. However, with multiple covariate index functions a useful dominance condition will require additional structure beyond the weak monotonicity used in the single index model above.

3 Conditional Moment Inequalities for Multinomial Choice

As in section 2, the data will be assumed to have a group structure, where i indexes the groups and t indexes observations within a group. There are a number of familiar multinomial choice applications with this group structure. In panel data applications in Labor and Public Finance, i typically indexes individuals, and t indexes time periods, though alternative groupings can also be relevant (an example from the study of hospital choice has i indexing illness categories and t indexing the individuals in these categories, see Ho and Pakes (forthcoming)). In Industrial Organization and Marketing applications, i would typically index markets and t would index either the different consumers in those markets (in demand analysis) or the firms that compete in them (in the analysis of a firm's choice of controls).

Observation (i, t) faces a number of choices. Each choice d has an associated random utility, $U_{d,i,t}$, and the observed choice, $y_{i,t}$, maximizes the random utility over choices. Take the number of choices to be $\mathscr{D}$ and number these choices so that $d = 1, \ldots, \mathscr{D}$. We consider the case of unordered response, where the numbering associated with each choice is arbitrary.[3] We could allow the set of choices to vary in an arbitrary way over i (as would be needed in most applications where i indexes markets) and obtain the same results as we present below, but to simplify the exposition we suffice with a constant choice set.

[3] Inequalities for models with ordered responses are considered in Pakes, Porter, Ho, and Ishii (forthcoming).

Given covariates $x_{i,t}$ for observation (i,t), the random utility for choice d takes the form

$$U_{d,i,t} = g_d(x_{i,t}, \theta_0) + f_d(\lambda_{d,i}, \varepsilon_{d,i,t}), \tag{2}$$

where $g_d(x_{i,t}, \theta_0)$ is a choice-specific function of observed characteristics, $x_{i,t}$. The term $\lambda_{d,i}$ denotes choice-specific effects which account for unobserved characteristics of choice d that do not vary across t. The term $\varepsilon_{d,i,t}$ represents any remaining unobserved, idiosyncratic determinants of the random utility. Notice that we require the index function to be additively separable here (we compare this to the weakly monotone transformation which does not have this requirement below). The literature we are aware of further restricts the $f_d(\cdot)$ in equation (2) to be additively separable in its arguments, that is

$$f_d(\lambda_{d,i}, \varepsilon_{d,i,t}) = \lambda_{d,i} + \varepsilon_{d,i,t}, \tag{3}$$

so we maintain this also, though it is not necessary for any of the results below.[4]

The observed choice, $y_{i,t}$, for agent (i,t) maximizes the random utility over choices:

$$y_{i,t} \in \operatorname*{argmax}_d U_{d,i,t}. \tag{4}$$

where the argmax function generates the set of choices that maximize random utility. If a single choice is the unique maximizer of random utility, then equation (4) determines the observed choice for (i,t). If there are multiple utility maximizing choices, then the argmax is a set consisting of the choices with maximal utility, and the agent can choose any element of the argmax set. The choices in this argmax set all have the same random utility value.

We refer to the case where different choices have the same random utility as "ties". Note that in the case of ties, the model in (4) remains agnostic about the mapping from the argmax set of choices to the observed choice, $y_{i,t}$. In particular, when we deal with the case of ties below we will not need to consider how the observed choice is selected from this group. It is this fact that enables us to generalize our model to account for set-valued regressors. However, because the case of ties is not essential to understanding the moment inequalities we derive, we start by assuming that the observed choice uniquely maximizes random utility.

The setup thus far is a standard random utility formulation of multinomial choice except that, as in Chamberlain (1980), we have allowed for a choice-specific group fixed effect. The covariates $x_{i,t}$ will need to vary by t to distinguish the index function from the fixed effect. The index function $g_d(x_{i,t}, \theta_0)$ is general enough to allow for the usual linear multinomial

[4] Strictly speaking, this further restriction on the unobservables entails no loss of generality. As noted in section 2, the fixed effect could be absorbed into the disturbance without loss of generality under the group homogeneity assumption.

logit functional form, ${x_{i,t}}'\theta_{0,d}$, where the parameter is partitioned by choice, and the usual conditional logit form, ${x_{d,i,t}}'\theta_0$, where the covariates differ by choice. The main parameter of interest will be θ_0.

Our key stochastic assumption is a generalization of Assumption M to allow for the multiple disturbances and multiple fixed effects corresponding to the multinomial choices. Notationally we let $\varepsilon_{i,t} = (\varepsilon_{1,i,t}, \ldots, \varepsilon_{\mathscr{D},i,t})'$ and $\lambda_i = (\lambda_{1,i}, \ldots, \lambda_{\mathscr{D},i})'$.

Assumption 1 *Given the conditioning set $(x_{i,s}, x_{i,t}, \lambda_i)$, for any $s \neq t$, the conditional distributions of $\varepsilon_{i,s}$ and $\varepsilon_{i,t}$ are the same:*

$$\varepsilon_{i,s} \big| x_{i,s}, x_{i,t}, \lambda_i \;\sim\; \varepsilon_{i,t} \big| x_{i,s}, x_{i,t}, \lambda_i.$$

As in section 2, Assumption 1 is a group homogeneity assumption on the disturbances. No parametric distributional restrictions are placed on the distribution of $\varepsilon_{i,t}$ (see section 5). Indeed, the distribution of these disturbances can have bounded or unbounded support and can have both mass points and continuous components. Perhaps more importantly the marginal distribution of the disturbances is allowed to vary arbitrarily across choices (d), and there are no restrictions on the covariance matrix of disturbances across choices. As a result, neither independence of irrelevant alternatives, nor any other limitation on the substitutability of different choices induced by the covariance structure of disturbances (such as the limited substitutability property discussed in Berry and Pakes 2007) is a source of concern. As a result, this specification nests both the familiar panel data model with individual choice-specific fixed effects and i.i.d. disturbances, a special case of which is Chamberlain's (1980) conditional logit model, and many differentiated product demand models for micro data (e.g. Berry, Levinsohn, and Pakes 2004).

We also note that the familiar panel data model assumption of strict exogeneity, that is

$$\varepsilon_{i,s} \big| x_{i,1}, \ldots, x_{i,T_i}, \lambda_i \;\sim\; \varepsilon_{i,t} \big| x_{i,1}, \ldots, x_{i,T_i}, \lambda_i,$$

is a special case of Assumption 1. Variants of strict exogeneity have long been used for identification of linear and nonlinear panel models, see Chamberlain (1982), Honore (1992), and Chernozhukov, Fernández-Val, Hahn, and Newey (2013). Additionally note that Assumption 1 does not restrict the covariances of the joint distribution of $(\epsilon_{i,s}, \epsilon_{i,t})$; all we require is that the marginal distributions of the two disturbance vectors be the same. So in the panel data context the disturbances for the different choices can be freely correlated across time. We come back to the issues of serial correlation and strict exogenity in our extensions, where we consider modifying Assumption 1 in different ways.

Assumption 1 does restrict the relationship between the disturbances and the covariates. For instance, heteroskedasticity would need to take a specific form where the heteroskedasticity in $\varepsilon_{i,t}$ is the same as $\varepsilon_{i,s}$ even when $x_{i,t} \neq x_{i,s}$. For example, if the heteroskedasticity in both $\varepsilon_{i,t}$ and $\varepsilon_{i,s}$ depended on $x_{i,t} + x_{i,s}$, then Assumption 1 would not be violated. One of our extensions (section 3.3) will allow some relaxation of this assumption when the heteroskedasticity takes on a known form. Of course, independence of disturbances and covariates across different s and t would suffice to satisfy Assumption 1.

By restricting the conditional joint distribution of the disturbances across the random utility choices to be the same for observations in group i, Assumption 1 enables us to learn about the relative response probabilities by comparing the observable components of random utilities across t for that group i. This within-group comparison does not depend on the joint distribution of disturbances across choices in any way. Moreover, though estimation and inference can combine the information on θ from different groups, the distribution of disturbances is allowed to vary in an arbitrary way across those groups.

3.1 Illustrative Moment Inequality

Given the random utility framework above along with Assumption 1, we can derive a set of moment inequality conditions that can be taken to data for inference on the parameter θ_0. We begin with a single conditional moment inequality that makes both the assumptions and logic underlying our conditional moment inequality analysis transparent. Following this derivation, we show how an extension of this logic leads to the complete set of conditional moment inequalities we use.

To simplify notation for this section of the paper, we eliminate the group i index with the understanding that all variables below are associated with the same group. We also assume that the probability of random utility "ties" is zero. That is, $\Pr(U_{c,t} = U_{d,t}) = 0$ for all $c \neq d$. We will explicitly include the case where ties can occur with any probability when our complete set of moment inequalities is derived below.

The probability that the choice by t, denoted by y_t, is equal to d is given by

$$\Pr(y_t = d | \Omega_t) = \Pr\left(\lambda_d + \varepsilon_{d,t} \ \geq \ \max_{c \neq d} \left\{ \Big[g_c(x_t, \theta_0) - g_d(x_t, \theta_0) \Big] \ + \ \lambda_c + \varepsilon_{c,t} \right\} \Bigg| \, \Omega_t \right), \qquad (5)$$

where Ω_t can be any conditioning set. This probability involves the difference of the index functions across the choices. Since we have assumed that the probability of "ties" is zero, the inequality in the above probability statement could be expressed equivalently as a strict inequality.

To establish the intuition behind our moment inequality conditions, consider the index function differences inside the probability in (5) above. In particular, suppose that if we compare the expression inside the square brackets in (5) for observations s and t, we find that for all $c \neq d$

$$g_c(x_t, \theta_0) - g_d(x_t, \theta_0) \; > \; g_c(x_s, \theta_0) - g_d(x_s, \theta_0). \tag{6}$$

That is, choice d maximizes the difference between observations s and t in the structural (or parameterized) determinants of the utility of the available choices when the structural part of the utility function is evaluated at $\theta = \theta_0$.

Then, taking the conditioning set in (5) to match the conditioning set in Assumption 1, we obtain

$$\Pr(y_s = d | x_s, x_t, \lambda) \; \geq \; \Pr(y_t = d | x_s, x_t, \lambda)$$

due to Assumption 1 and the observation that s and t share the same choice-specific fixed effects as members of the same group.

Notice that the inequality in (6) can be checked by the analyst for any θ. Also, this condition can be re-arranged so that equation (6) holds for all $c \neq d$ if and only if

$$g_d(x_s, \theta_0) - g_d(x_t, \theta_0) \; > \; g_c(x_s, \theta_0) - g_c(x_t, \theta_0). \tag{7}$$

Consequently

$$d = argmax_c\Big(g_c(x_s, \theta_0) - g_c(x_t, \theta_0)\Big) \;\; \Rightarrow \;\; \Pr(y_s = d | x_s, x_t, \lambda) \; \geq \; \Pr(y_t = d | x_s, x_t, \lambda). \tag{8}$$

Moreover since (8) holds for every observation couple within each group it will hold for averages across couples in different groups even if there are only two members in each group and the disturbance distribution and choice-specific fixed effect vary arbitrarily across groups. It is these facts that underlie our moment inequality estimators.

Equation (7) considers the differences in the covariate index between t and s and labels the choice corresponding to the maximum difference d. So, the corresponding stochastic ordering on the conditional outcome probabilities in (8) is only dependent on the magnitude of the covariate index differences. In particular this ordering does not depend on the disturbances, and hence does not induce a selection problem. This is where the additive separability of $g_d(x_t, \theta_0)$ in the random utility from $(\lambda_d, \varepsilon_t)$ is used in our derivations. In the single covariate index case of section 2, only weak monotonicity in the covariate index was used and separability was not needed. In contrast, with multiple covariate indices for multinomial choice, the magnitude of covariate index changes is needed to establish which outcome can be stochastically ordered.

3.2 Ties

Before deriving the complete set of moment inequalities, we reintroduce the possibility of random utility ties. We note, however, that allowing for ties is not essential to the main argument that leads to our conditional moment inequalites. So a reader who does not want to focus on the added detail that accompanies our treatment of ties can skip this subsection. For that reader, we specialize our moment inequality result to the case without ties in the next subsection.

There are two main reasons for allowing random utility ties in our framework. The generality of Assumption 1 allows for the distributions of disturbances and covariates to include mass points, which then implies that ties in random utility could occur with positive probability. Often discrete choice models will include assumptions that force the probability of random utility ties to be zero. In contrast, our framework allows for ties and yet imposes no structure on the relationship of the *observed* choice to the set of utility maximizing choices. That is, we place no restrictions on the rule that selects the observed choice from among the equally-valued utility-maximizing choices. Second, as we will show in the extensions to our basic result, allowing for ties enables us to apply our findings to cases where there are set-valued regressors. As noted there, the set-valued regressor results allows us to handle several problems which appear quite frequently in applications of discrete choice modeling.

As above, we will forgo the i subscript and note that every variable stated below corresponds to a given group i. Consider the choice problem for t. If the random utility for a choice d is the *unique* maximizer of random utilities, then d is clearly the choice: $y_t = d$. If the random utility for choice d is one of multiple maximizers, then d is among the possible choices that could be observed. So, letting $U_t = \{U_{1,t}, \ldots, U_{\mathscr{D},t}\}$,

$$\{U_t : U_{d,t} > \max_{c \neq d} U_{c,t}\} \subseteq \{U_t : y_t = d\} \subseteq \{U_t : U_{d,t} \geq \max_{c \neq d} U_{c,t}\}. \tag{9}$$

The first set of U_t vectors generate choice d as the unique maximizer of random utility (so this condition is sufficient for d to be chosen), and the last set of U_t vectors generate choice d being among the set of possible maximizers (this condition is necessary for d to be chosen). The set relations come from noting that if d is the unique maximizer then y_t takes the value of d. On the other hand, if y_t takes the value of d, then d must be included in the set of random utility maximizing choices. In the special case where "ties" cannot occur, there is a unique maximizer, and the three sets are identical.

When random utilities are handled in this way, the choice model is formally incomplete (Tamer 2003). In particular, the distribution of random utilities (as determined by the distribution of covariates, fixed effects, and disturbances) need not fully determine the probability of choices. As noted by Tamer (2003) in a multiple agent context, even when there is not a uniquely determined choice or outcome, the necessary conditions for a choice to be made may still lead to inequalities on the probabilities of various choices that then provide information on the unknown parameters. The results below show that this can also occur in a standard discrete choice problem, and, when incompleteness is allowed, the set relationships in equation (9) imply conditional moment inequalities that do not depend on the distribution for the disturbance vector.[5]

In addition to random utility ties, there is another source of potential "ties". Notice that in equations (6) - (8), we consider the case where there is a unique choice that maximizes the index function differences. More generally, we will now also allow for the possibility that index function differences could be equal for different choices. Ties of this kind could come from discreteness in the covariates, or, when evaluating the index function differences at various values of θ, one could consider a parameter value that equates index differences.

3.3 Implied Moment Inequalities

The probability inequality in (8) is based on the choice that maximizes the difference of index functions. We can push this logic further to obtain similarly motivated inequalities based on a rank ordering of the index function differences across the choices. For a pair of decisions s and t, start by ordering the difference of index functions by choice. Without ties, there's a unique value of the difference $g_d(x_s, \theta) - g_d(x_t, \theta)$ for each d. Allowing for ties, let $K(x_s, x_t, \theta)$ denote the number of distinct values of the difference $g_d(x_s, \theta) - g_d(x_t, \theta)$ among the choices $d = 1, \ldots, \mathscr{D}$. So, $1 \leq K(x_s, x_t, \theta) \leq \mathscr{D}$, and, when we order the index function differences, there are $K(x_s, x_t, \theta)$ distinct rank values.

Given a value of θ, let the choices corresponding to the minimum difference of index functions be denoted

$$D^{(1)}(x_s, x_t, \theta) = \underset{c \in \{1, \ldots, \mathscr{D}\}}{\arg\min} \Big[g_c(x_s, \theta) - g_c(x_t, \theta) \Big].$$

The set of choices with the largest index function differences will be denoted

$$D^{(K(x_s,x_t,\theta))}(x_s, x_t, \theta) = \underset{c \in \{1, \ldots, \mathscr{D}\}}{\arg\max} \Big[g_c(x_s, \theta) - g_c(x_t, \theta) \Big].$$

$D^{(1)}(x_s, x_t, \theta)$ will contain a single choice if there is a unique minimizer of the index function differences at θ and multiple choices if there are a set of minimizers. Since all the choices

[5] An alternative way to handle ties would be to assume a known selection mechanism in the case of ties, analogous to an equilibrium selection mechanism in the games considered by Tamer (2003).

contained in $D^{(1)}(x_s, x_t, \theta)$ have the same index function difference, we refer to each such set of choices as an equivalence set of choices.

These ordered equivalence sets are formally defined as follows. Suppose $w, v \in \{1, \ldots, K(x_s, x_t, \theta)\}$. For any $c, d \in D^{(w)}(x_s, x_t, \theta)$, $g_c(x_s, \theta) - g_c(x_t, \theta) = g_d(x_s, \theta) - g_d(x_t, \theta)$. If $v < w$, then for any $c \in D^{(v)}(x_s, x_t, \theta)$ and $d \in D^{(w)}(x_s, x_t, \theta)$,

$$g_c(x_s, \theta) - g_c(x_t, \theta) < g_d(x_s, \theta) - g_d(x_t, \theta), \tag{10}$$

so that there is a strict inequality between index differences for choices in different equivalence classes. At times it will be convenient to use more compact notation, and let $K_{s,t}(\theta) \equiv K(x_s, x_t, \theta)$ while $K_{s,t} \equiv K_{s,t}(\theta_0)$. Similarly, we let $D^{(w)}_{s,t}(\theta) \equiv D^{(w)}(x_s, x_t, \theta)$ and $D^{(w)}_{s,t} \equiv D^{(w)}_{s,t}(\theta_0)$.

The choices have now been partitioned into index function difference equivalence sets. The ranks of these equivalence sets generate the desired results on relative conditional probabilities. For instance, we can directly extend the result in (8) to conclude that

$$D^{(K_{s,t})}_{s,t} = \underset{c \in \{1,\ldots,\mathscr{D}\}}{\arg\max} \Big(g_c(x_s, \theta_0) - g_c(x_t, \theta_0) \Big)$$

implies

$$\Pr\Big(y_s \in D^{(K_{s,t})}_{s,t} \big| x_s, x_t, \lambda \Big) \ \geq \ \Pr\Big(y_t \in D^{(K_{s,t})}_{s,t} \big| x_s, x_t, \lambda \Big), \tag{11}$$

and now the inequality allows for random utility ties. By accounting for ties in the random utilities, the distribution of $\varepsilon_{i,t}$ is allowed to have mass points or have bounded support. Similarly, the distribution of $x_{i,t}$ is also unrestricted.

To obtain the inequality (11), we first extend the set relations given in (9) to subsets of $\mathscr{D}$. Let D denote any non-empty set of choices. We will use the following relationships

$$\bigcup_{d \in D} \left\{ U_t : U_{d,s} > \max_{c \notin D} U_{c,s} \right\} \ \subseteq \ \{U_t : y_t \in D\} \ \subseteq \ \bigcup_{d \in D} \left\{ U_t : U_{d,t} \geq \max_{c \notin D} U_{c,t} \right\}. \tag{12}$$

We now derive the probability inequality in (11) for the highest ranked equivalence set. Since the derivation will also suffice for additional cases to be introduced below, we let $D = D^{(K_{s,t})}_{s,t}$ and $\Omega_{s,t} = \{x_s, x_t, \lambda\}$. Then we can re-define D and $\Omega_{s,t}$ to cover other cases of interest. Finally, we have

$$\begin{aligned} \Pr\left(y_s \in D|\Omega_{s,t}\right) &\geq \Pr\left(\bigcup_{d\in D}\left\{U_s : U_{d,s} > \max_{c\notin D}\ U_{c,s}\right\}\Big|\Omega_{s,t}\right) \qquad (13)\\ &= \Pr\left(\bigcup_{d\in D}\left\{\lambda_d + \varepsilon_{d,s} > \max_{c\notin D}\ \left(\Big[g_c(x_s,\theta_0) - g_d(x_s,\theta_0)\Big] + \lambda_c + \varepsilon_{c,s}\right)\right\}\Big|\Omega_{s,t}\right)\\ &\geq \Pr\left(\bigcup_{d\in D}\left\{\lambda_d + \varepsilon_{d,s} \geq \max_{c\notin D}\ \left(\Big[g_c(x_t,\theta_0) - g_d(x_t,\theta_0)\Big] + \lambda_c + \varepsilon_{c,s}\right)\right\}\Big|\Omega_{s,t}\right)\\ &= \Pr\left(\bigcup_{d\in D}\left\{\lambda_d + \varepsilon_{d,t} \geq \max_{c\notin D}\ \left(\Big[g_c(x_t,\theta_0) - g_d(x_t,\theta_0)\Big] + \lambda_c + \varepsilon_{c,t}\right)\right\}\Big|\Omega_{s,t}\right)\\ &= \Pr\left(\bigcup_{d\in D}\left\{U_t : U_{d,t} \geq \max_{c\notin D}\ U_{c,t}\right\}\Big|\Omega_{s,t}\right)\\ &\geq \Pr\left(y_t \in D|\Omega_{s,t}\right). \end{aligned}$$

The first and last inequalities follow by the set inclusions in (12). The second equality follows by Assumption 1. Note that the second inequality is a weak inequality but it uses the fact that for $d \in D^{(K_{s,t})}$ and $c \notin D^{(K_{s,t})}$, equation (10) gives the strict inequality $g_c(x_t,\theta_0)-g_d(x_t,\theta_0) > g_c(x_s,\theta_0) - g_d(x_s,\theta_0)$.

To obtain the probability inequality in (13), a comparison of random utilities is made between choices in $D^{(K_{s,t})}$ and in the complementary set $D^{(K_{s,t}-1)} \cup \ldots \cup D^{(1)}$. An analogous set of comparisons can be made after redefining

$$D \equiv D^{(K_{s,t})} \cup \ldots \cup D^{(K_{s,t}-w)},$$

and the complementary set of choices becomes

$$D^{(K_{s,t}-w-1)} \cup \ldots \cup D^{(1)}, \text{ for } w = 0, \ldots, \mathscr{D} - 2.$$

In fact, setting $D = D^{(K_{s,t})} \cup \ldots \cup D^{(K_{s,t}-w)}$, then the derivation in (13) yields

$$\Pr\left(y_s \in D^{(K_{s,t})} \cup \ldots \cup D^{(K_{s,t}-w)}|\Omega_{s,t}\right) \geq \Pr\left(y_t \in D^{(K_{s,t})} \cup \ldots \cup D^{(K_{s,t}-w)}|\Omega_{s,t}\right)$$

for any for $w = 0, \ldots, \mathscr{D} - 2$.

This set of probability inequalities leads directly to a set of corresponding conditional moment inequalities which are stated formally in the proposition below. Define the moment functions

$$m_w(y_s, y_t, x_s, x_t, \theta) = \mathbf{1}\left\{y_t \in \bigcup_{r=0}^{w}\{D^{(K(x_s,x_t,\theta)-r)}(x_s, x_t, \theta)\}\right\} - \mathbf{1}\left\{y_s \in \bigcup_{r=0}^{w}\{D^{(K(x_s,x_t,\theta)-r)}(x_s, x_t, \theta)\}\right\}$$

for $w = 0, \ldots, K(x_s, x_t, \theta) - 2$, and reintroduce the i subscript to be clear about the dependence on the group structure.

Proposition 2 *For any set of observations $(i,t)_{t=1}^{T_i}$ making choices by maximizing (2), if Assumption 1 is satisfied then, for $s \neq t$,*

$$0 \ \leq \ E\Big[m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta_0) \ \big| \ x_{i,s}, x_{i,t}\Big]$$

for $w = 0, 1, \ldots, K(x_{i,s}, x_{i,t}, \theta_0) - 2$, a.s. $(x_{i,s}, x_{i,t})$.

The proposition is obtained by first conditioning on $(x_{i,s}, x_{i,t}, \lambda_i)$ and then integrating out with respect to the distribution of λ_i conditional on $(x_{i,s}, x_{i,t})$ in order to formulate the inequalities in terms of observable conditioning sets. The usefulness of these moment inequalities in applications will come from corresponding inequalities for their empirical analogues evaluated at values of θ. Note that these moment inequalities may be consistent with other values of θ (as well as with θ_0), a point which we discuss further below.

Notice that when $x_{i,s} = x_{i,t}$, all choices are in the same equivalence class regardless of value of θ, and hence the moment functions are identically zero for all θ. Similarly, if $y_{i,s} = y_{i,t}$ then $m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta) = 0$ for all w and θ. In either of these cases, this pair of observations would not provide information about the true value θ_0 through an empirical analog of a moment inequality derived from Proposition 2.

Now consider the case where $y_{i,s} \neq y_{i,t}$ and $x_{i,s} \neq x_{i,t}$, and assume that as θ changes so does the ordering of at least some of the index differences $g_d(x_{i,s}, \theta) - g_d(x_{i,t}, \theta)$. For simplicity, consider the case where there are no ties in the index function differences, so that there is a unique choice associated with each rank. For a fixed θ, the number of possible non-zero conditional moment functions is equal to the difference in ranks of the choices corresponding to $y_{i,s}$ and $y_{i,t}$. If the rank of $y_{i,s}$ for a particular value of θ is larger than the rank of $y_{i,t}$, then all these possible non-zero conditional moment functions are positive (and equal to one). This combination of observations and θ provides evidence that is consistent with the moment inequalities in Proposition 2. If, on the other hand, the rank of $y_{i,t}$ is greater than $y_{i,s}$, these possible non-zero conditional moment functions will be negative, providing evidence against this θ being the true value of the parameter. In this case the larger the difference in ranks, the greater the number of negative conditional moments, and the greater the evidence against that parameter value.

It might also be instructive to compare using the complete set of conditional moment inequalities in Proposition 2 to what would happen if we focused only on the single conditional moment inequality generated by the highest ranked index difference. The single conditional moment inequality from only considering the largest index difference is the "illustrative moment inequality" of section 3.1 given in (8). This single conditional moment inequality is equivalent to using only the first conditional moment, m_0, in Proposition 2. Now consider the information about θ_0 contained in a pair of observations when the analysis only employs this one moment. The conditional moment function, m_0, would only be non-zero if exactly one of $y_{i,s}$ or $y_{i,t}$ corresponds to the highest ranked choice for a given θ (so if the second ranked difference were chosen we would not use the information in that comparison even if the second ranked difference was very close in value to the first ranked difference for that θ). When there are a large number of choices, a non-zero conditional "m_0" moment function is unlikely to occur, regardless of the θ value at which it is evaluated. By using the conditional moment inequalities for the full range of choice rankings, we are able to glean considerably more information about θ_0 from each individual pair of observations.

Limits and Incidental Parameters. When considering the asymptotic properties of estimation and inference procedures for this problem, the limits could be taken as either n, T_i, or both grow large. When the appropriate limit has n growing large the model has the usual fixed effects "incidental parameter" problem. This problem is circumvented by our method of within-group differencing of the index functions, which exploits the separability of the fixed effects and covariate index functions in the random utility specification. Chamberlain (1980) considers the same problem in a conditional logit model, and uses the form of the conditional likelihood to eliminate additive fixed effects. The main difference between the two frameworks is that we allow for a free joint distribution of choice-specific disturbances but must suffice with partial (instead of point) identification. Examples with a large number of individuals (n) observed over a short time period (T_i) are familiar from panel data applications. An extreme case of T_i large and n small occurs when the data is cross sectional and is considered a single group. Then $n = 1$ and the λ are choice specific constants. Consumer demand models are often similar in that there frequently are many observations per market (T_i) but only a small number of markets (n). Cases where T_i and n are of approximately equal size often occur in marketing problems when samples are drawn from a large number of cities, and in game theoretic equilibrium problems in I.O. where the number of agents and the number of markets are often of a similar magnitude. To formally analyze the limiting properties of particular estimators, one would need to specify the dependence structure as the sample grows in either (or both) dimension(s). We leave that development to future research and application of these methods.

Identified Set. If $\Theta_{0,n}$ is defined as the set of parameters that satisfy the conditional moment inequalities in Proposition 2,

$$\Theta_{0,n} = \left\{ \theta \in \Theta : \bigcap_{i=1}^{n} \bigcap_{\substack{s,t=1 \\ s<t}}^{T_i} \bigcap_{w=0}^{K(x_{i,s},x_{i,t},\theta)-2} E\Big[m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta) \;\Big|\; x_{i,s}, x_{i,t} \Big] \geq 0 \;\text{ a.s. } \{(x_{i,s}, x_{i,t})\} \right\},$$

then the content of the proposition is that $\theta_0 \in \Theta_{0,n}$.

We have stated Proposition 2 and defined $\Theta_{0,n}$ to hold almost surely over the covariates. We could have made the same statements conditional on the covariate values in the sample. Conditioning only on the observed values would have a repeated sampling interpretation over these values, and justifies examining the structure of $\Theta_{0,n}$ without making additional assumptions on the data generating process.[6] Either way the implied $\Theta_{0,n}$ is a finite sample object that can change with sample size. Note that $\forall n$, $\Theta_{0,n+1} \subseteq \Theta_{0,n}$, so any limit of the identified set will be a subset of $\Theta_{0,n}$ and yet will still contain θ_0.

A sufficient condition for a value of θ to be in $\Theta_{0,n}$ is that, with probability one, θ and θ_0 generate exactly the same ranking of index function differences.[7] Though this condition is sufficient to insure $\theta \in \Theta_{0,n}$, it is not necessary. That is, there can be a θ which does not preserve the same index function difference rankings as θ_0 that is in $\Theta_{0,n}$. For example, assume that when evaluated at $\theta = \theta_0$, d_a is the choice that was ranked highest (maximal) when we consider the differences in the choice indexes between observations (i, s) and (i, t), and let d_b be the second highest ranked difference. Now consider a $\theta^* \neq \theta_0$ which reverses

[6] Under the almost sure definition of $\Theta_{0,n}$ if the observable random variables are identically distributed across i, then the intersection over i in the definition above is redundant or unnecessary. Similarly if the random variables are identically distributed across t for each group i then the intersection over s and t subscripts is redundant. That is, the set would be well defined for a single pair $s < t$. An alternative definition would be to first set

$$\Theta_{0,i,s,t} = \left\{ \theta \in \Theta : \bigcap_{w=0}^{K(x_{i,s},x_{i,t},\theta)-2} E\Big[m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta) \;\Big|\; x_{i,s}, x_{i,t} \Big] \geq 0 \;\text{ a.s. } (x_{i,s}, x_{i,t}) \right\}$$

and then consider the identified set defined by the intersection of these sets across i and t. The distinction from the definition above would come from the support set of the joint distribution of the conditioning sets $(x_{i,s}, x_{i,t})$. In particular, $\Theta_{0,n} \subseteq \cap_{i=1}^{n} \cap_{\substack{s,t=1 \\ s<t}}^{T_i} \Theta_{0,i,s,t}$.

[7] Since applying an affine transformation to all choice-specific index functions within any group will leave these ranks unchanged our ability to differentiate between different θ values will, at best, be up to an affine transform of the index functions (as is true in linear parametric discrete choice models).

these two rankings but leaves all other rankings unchanged. Provided the probability of d_b for observation (i,s) is higher than the probability of d_b for individual (i,t), the identified set will include that θ^*. On the other hand, failure of this kind of probability inequality when index function difference rankings differ at any $(x_{i,t}, x_{i,s})$ with positive measure will ensure that $\theta^* \notin \Theta_{0,n}$.

Proposition 2 and Manski (1987) Proposition 2 can be considered an extension of the Manski (1987) approach for panel binary choice to multinomial choice problems with a general index function. To demonstrate the subtle differences between the current work and Manski (1987), we specialize Proposition 2 to the linear binary choice case (i.e. $\mathscr{D} = 2$ and, less important to the argument, $g_d(x,\theta) = {x_d}'\theta$), which was the case considered in Manski (1987). Then Proposition 2 becomes

$$(x_{2,i,s} - x_{2,i,t})'\theta_0 \,>\, (x_{1,i,s} - x_{1,i,t})'\theta_0 \quad\Rightarrow\quad \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) \geq \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t}). \tag{14}$$

Manski (1987) also assumes that the support of $\varepsilon_{2,i,t} - \varepsilon_{1,i,t} \,|\, x_{i,s}, x_{i,t}, \lambda_i$ is the real line. With the binary analog of Assumption 1 and this additional assumption, Manski (1987) obtains the stronger conclusion that

$$\begin{aligned}
&(x_{2,i,s} - x_{2,i,t})'\theta_0 \,>\, (x_{1,i,s} - x_{1,i,t})'\theta_0 \quad\Leftrightarrow\quad \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) > \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t})\\
&\text{and}\\
&(x_{2,i,s} - x_{2,i,t})'\theta_0 \,=\, (x_{1,i,s} - x_{1,i,t})'\theta_0 \quad\Leftrightarrow\quad \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) = \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t}).
\end{aligned} \tag{15}$$

Recall that the general result from the current paper (equation (14)) provides a *sufficient* condition for the difference in probabilities to be non-negative. The result in equation (15) provides *necessary and sufficient* conditions for the sign of the difference in probabilities. Manski (1987) shows that these necessary and sufficient conditions lead to a straightforward argument for point identification. For multinomial choice, our result states, for instance, that if d is the choice corresponding to the highest ranked index difference, then $Pr(y_s = d|x_s, x_t) \geq Pr(y_t = d|x_s, x_t)$. However, if there is a choice a, such that $Pr(y_s = a|x_s, x_t) \geq Pr(y_t = a|x_s, x_t)$, then clearly choice a need not correspond to the largest index function difference. In fact, in general, there will be many such choices a (with $Pr(y_s = a|x_s, x_t) \geq Pr(y_t = a|x_s, x_t)$) and all of these choices cannot simultaneously correspond to the largest index function difference. As a result, it is not *necessary* for d to be the highest ranked choice to have $Pr(y_s = d|x_s, x_t) \geq Pr(y_t = d|x_s, x_t)$.

The other minor difference between the approach in this paper and Manski (1987) is in the treatment of ties. Manski (1987) makes a fairly standard binary choice assumption that $y_{i,t} = 2$ if $U_{2,i,t} \geq U_{1,i,t}$. Implicitly, this assumption provides a particular selection rule to deal with ties (random utility ties necessarily lead to observing choice 2). The results of this paper do not require specification of a selection rule for ties, a point which is integral to the extensions in the next section.

Computation and Estimation A computational advantage of basing estimation on moments derived from Proposition 2 is that it would not require the estimation of choice probabilities at different parameter values. This is typically the computationally costly step in estimating multinomial choice models with parametric distributions of disturbances. The estimation algorithms for the models with an assumed parametric disturbance distribution base their objective functions on the difference between the observed outcomes and the model's predicted probabilities at different values of $\theta \in \Theta$. In parametric problems, those predicted probabilities are especially computationally burdensome when (i) there are many choices; and/or (ii) the joint distribution of disturbances has a rich pattern of dependence across choices. In contrast, ranking the index function differences is a straightforward calculation involving only a sort (or ranking) algorithm, and the degree of computational difficulty has no relationship at all to the covariance structure of the disturbances from the choices.

Of course the rankings are inherently discontinuous in θ and this may impose an additional computational burden on the search algorithm (a problem related to that which occurs often in moment inequality models, and in parametric discrete choice problems which use frequency simulators).[8] Indeed, since the empirical analog of the moments we use are straightforward to calculate, the main computational cost of our approach will be in finding the estimate of the identified set and characterizing the aspects of its distribution needed for inference. The computational burden of these steps is likely to vary directly with the dimension of the parameter space. For a discussion of these issues, see Bar and Molinari (2013).

The literature on inference on θ_0 based on conditional moment inequalities is new and developing. Andrews and Shi (2013) proposes a method of generating a set of unconditional moment inequalities that provide asymptotically equivalent inference to the conditional moment inequalities. In principle, generating unconditional moments from the conditional moments requires choosing positive functions of the conditioning variables as "instruments." Andrews and Shi (2013) show how to make these choices systematically to generate the desired equivalence. We note that there are some natural positive instrument functions for

[8] We leave open the question of whether one could do better with smoothed rankings, as in Yan (2013) for future research.

use based on equation (10). In particular, the difference of the index functions are already ranked for each θ.[9] Other methods for conditional moment inequality inference could also be employed (see, for e.g., Armstrong 2011, Chetverikov 2011, Chernozhukov, Lee, and Rosen 2013, and Aradillas-López, Gandhi, and Quint 2013).

4 Extensions

We now extend the framework for developing conditional moment inequalities for partial identification that was presented in the previous section. The extensions we consider enable the analysis of discrete choice models with: (i) some forms of dependence of the disturbance distribution on the regressors (or of "endogeneity"), and (ii) set-valued regressors.

4.1 Covariate Dependent Disturbance Distributions

This section considers modifications to Assumption 1 that allow us to accommodate different forms of endogeneity into our analysis of multinomial choice. We focus on two familiar cases which, in reverse order, are: (i) conditional heteroscedasticity, where we consider a generalization of Assumption 1, and (ii) regressors which depend on lagged values of the disturbance terms, where we need more restrictions on the conditional distribution of the disturbance term than those given in Assumption 1. This latter case includes models which have (one or more) lagged dependent variables as regressors. Since we continue to allow for choice-specific fixed effects, when there are lagged dependent variables the additional restrictions allow us to analyze panel data discrete choice models with both unobserved heterogeneity in preferences and state dependence, a problem which has been salient in analyzing several empirical issues.

4.1.1 Lagged Dependent Variables.

We modify the random utility model (2) to explicitly allow for dependence on the previous period's choice,[10] or on $y_{i,t-1}$, so

$$U_{d,i,t} = g_d(x_{i,t}, y_{i,t-1}, \theta_0) + f_d(\lambda_{d,i}, \varepsilon_{d,i,t}). \tag{16}$$

[9] For example, a natural choice for an instrument to interact with the difference in indicator functions leading to $m_w(y_s, y_t, x_s, x_t, \theta)$ would be $g_c(x_s, \theta) - g_c(x_t, \theta) - [g_d(x_s, \theta) - g_d(x_t, \theta)]$ where $c \in D^{(K_{s,t}(\theta))}(\theta)$ and $d \in D^{(K_{s,t}(\theta)-w)}(\theta)$.

[10] Here we refer to "periods" rather than "observations" as lagged dependent variable models typically have t denoting time.

A familiar special case, which has appeared repeatedly in the literature, is

$$U_{d,i,t} = g_d(x_{i,t}, \theta_0) + \{y_{i,t-1} = d\}\gamma_0 + \lambda_{d,i} + \epsilon_{d,i,t},$$

where γ_0 captures "state dependence" and the $\{\lambda_{d,i}\}$ capture state specific "heterogeneity."[11]

The issue of separating the impact of state dependence from that of heterogeneity has been discussed extensively in the econometrics literature (e.g. Heckman 1981) as it has been central to analyzing several empirical issues. Perhaps the oldest of these is the determination of unemployment durations (see the review by Kiefer (1998) and the more recent empirical work in Kroft, Lange, and Notowidigdo (2013)). More recently the same problem has been a focus of panel data demand models in which there is a need to separate the impact of "switching costs" from that of unobserved heterogeneity (see, for example, Handel (2013)'s analysis of the choice of health insurance plans).

The model in equation (16) implies that $y_{i,t-1}$ will depend on past values of the disturbance vector. As a result, if the lagged dependent variable $y_{i,t-1}$ is included in the conditioning set of Assumption 1, then that group homogeneity assumption will, in general, be violated. This violation would, in turn, invalidate the proof of Proposition 2 given in (13).

To separate the influence of a lagged dependent variable and the fixed effect on current outcomes, we strengthen our group homogeneity assumption to rule out dependence across time in the joint distribution of $\varepsilon_{i,1}, \ldots, \varepsilon_{i,T_i}$. Let $\mathcal{J}_{i,t}$ denote the history of covariates and disturbances through period t for i: $\mathcal{J}_{i,t} \equiv \{x_{i,t}, x_{i,t-1}, \ldots, \varepsilon_{i,t}, \varepsilon_{i,t-1}, \ldots\}$. We could include the fixed effect λ_i in the definition of the set $\mathcal{J}_{i,t}$, but we will keep it separate to be clear about its influence.

Assumption L *For any t, the disturbance $\varepsilon_{i,t}$ is: (i) conditionally independent of current covariates $x_{i,t}$ and the history $\mathcal{J}_{i,t-1}$, and (ii) stationary across time, or*

$$\varepsilon_{i,t}|x_{i,t}, \mathcal{J}_{i,t-1}, \lambda_i \;\sim\; \varepsilon_{i,t}|\lambda_i \;\sim\; \varepsilon_{i,1}|\lambda_i. \quad \spadesuit$$

The content of Assumption L is that: (i) all the dependence over time in the unobservable determinants of the value of the various choices is picked up by the choice-specific fixed effects (λ_i), and (ii) conditional on the fixed effects, the distribution of $\varepsilon_{i,t}$ is constant over time. Notice, however, that $x_{i,t+\tau}$ for $\tau > 0$ can depend on $\varepsilon_{i,t}$; that is the other covariates can be endogenous in the same way that lagged choices are. In this sense, Assumption L is weaker than the strict exogeneity assumption discussed in Section 3 (a point we return

[11] Generalizations of the random utility model that allow for more than one lagged value of the choice to enter the index function, or for different lagged choices to enter the utility function of a given choice with different coefficients, can be analyzed analogously.

to below). Moreover, as before, Assumption L places no restrictions on the correlation of disturbances across choices.

We now modify the argument in (13) to accommodate Assumption L. As in section 3.3, index function differences can be ordered across choices. The only additional wrinkle is that these index functions now explicitly depend on a lagged dependent variable. We could formally re-define the choice equivalence sets for the lagged dependent case, but there should be no confusion in simply adopting the previous notation. Let D represent a choice equivalence set based on index function differences, as stated prior to (13), and suppose $s > t$. Then

$$\begin{aligned}
\Pr\left(y_s \in D | x_s, \mathcal{J}_{s-1}, \lambda\right) \;&\geq\; \Pr\left(\bigcup_{d\in D}\left\{U_s : U_{d,s} > \max_{c\notin D}\; U_{c,s}\right\}\middle| x_s, \mathcal{J}_{s-1}, \lambda\right) \qquad (17)\\
=\;& \Pr\left(\bigcup_{d\in D}\left\{\lambda_d + \varepsilon_{d,s} > \max_{c\notin D}\;\left(\left[g_c(x_s, y_{s-1}, \theta_0) - g_d(x_s, y_{s-1}, \theta_0)\right] + \lambda_c + \varepsilon_{c,s}\right)\right\}\middle| x_s, \mathcal{J}_{s-1}, \lambda\right)\\
\geq\;& \Pr\left(\bigcup_{d\in D}\left\{\lambda_d + \varepsilon_{d,t} \geq \max_{c\notin D}\;\left(\left[g_c(x_t, y_{t-1}, \theta_0) - g_d(x_t, y_{t-1}, \theta_0)\right] + \lambda_c + \varepsilon_{c,t}\right)\right\}\middle| x_t, \mathcal{J}_{t-1}, \lambda\right)\\
=\;& \Pr\left(\bigcup_{d\in D}\left\{U_t : U_{d,t} \geq \max_{c\notin D}\; U_{c,t}\right\}\middle| x_t, \mathcal{J}_{t-1}, \lambda\right)\\
\geq\;& \Pr\left(y_t \in D | x_t, \mathcal{J}_{t-1}, \lambda\right),
\end{aligned}$$

where the second inequality follows by Assumption L.

There is an important distinction between the arguments in (13) and (17). Notice that the conditioning set never changes in (13), and hence the implied conditional moment inequalities are just equivalent restatements of the derived conditional probability inequalities. However, (17) shows $\Pr\left(y_s \in D | x_s, \mathcal{J}_{s-1}, \lambda\right) \geq \Pr\left(y_t \in D | x_t, \mathcal{J}_{t-1}, \lambda\right)$, where the conditioning set changes along with the outcome variable. To form the proposition that underlies our estimators we need to restate this inequality in terms of the expectation of a *difference* of indicator functions conditional on a single conditioning set. We are able to obtain such a result due to the particular structure of the two information sets, specifically $\{x_t, \mathcal{J}_{t-1}, \lambda\} \subset \{x_s, \mathcal{J}_{s-1}, \lambda\}$. This structure allows application of the law of iterated expectations to obtain a conditional moment inequality based on the earliest (smallest) information set. More formally we have the following proposition.[12]

[12] Note that the multinomial choice lagged dependent variable case includes the special case of binary choice with lagged dependent variables (and fixed effects). When the index function g_d takes a linear form, one can show that modifications of Manski (1987)'s assumptions (to include the lagged dependent variable) enable point identification of the parameter vector.

Proposition 2′ *For any individual (i,t) making choices by maximizing (16), if Assumption L is satisfied and $s > t$, then*

$$0 \ \leq \ E\Big[m_w(y_{i,s}, y_{i,s-1}, y_{i,t}, y_{i,t-1}, x_{i,s}, x_{i,t}, \theta_0) \ \big| \ x_{i,t}, x_{i,t-1}, \ldots, y_{i,t-1}, y_{i,t-2}, \ldots\Big]$$

for $w = 0, 1, \ldots, K_{s,t}(\theta_0) - 2$, a.s. $(x_{i,t}, x_{i,t-1}, \ldots, y_{i,t-1}, y_{i,t-2}, \ldots)$. ♠

The conditioning set in the conclusion of Proposition 2′ is important because it is the basis for the construction of instruments used for estimation and inference. Under Assumption L, x_t and lags are appropriate instruments, but x_s may not be a valid instrument (or any $x_{t+\tau}$ with $\tau > 0$). In applications where it is appropriate to strengthen Assumption L to a form that includes strict exogeneity with respect to the covariates:

$$\varepsilon_{i,t}|x_{i,T_i}, x_{i,T_i-1}, \ldots, x_{i,t}, \mathcal{J}_{i,t-1}, \lambda_i \ \ \sim \ \ \varepsilon_{i,t}|\lambda_i \ \ \sim \ \ \varepsilon_{i,1}|\lambda_i,$$

then we could strengthen Proposition 2′ to include in its conditioning set the leads and lags of the x's, so that both $x_{i,s}$ and $x_{i,t}$ would be available as instruments. Of course the cost of maintaining the strengthened assumption is that it does not allow the other covariates to be endogenous in the sense that the lagged values of the choice are.

4.1.2 Conditional Heteroskedasticity.

This section weakens Assumption 1 by assuming that group homogeneity holds only for a particular observable subset of the data. We illustrate with a particular conditional heteroskedasticity example of the form $\varepsilon_{d,i,t} = \varepsilon^*_{d,i,t}\sigma_d(v_{i,t})$, where $v_{i,t}$ is observed and the functional form of σ_d can be unknown. Any dependence of the distribution of $v_{i,t}$ on $x_{i,t}$ will typically violate Assumption 1. Note that this violation will occur even if $\varepsilon^*_{d,i,t}$ is independent of the $(x_{i,t}, v_{i,t})$ couple. On the other hand if we only compare observations which have the same value of $v_{i,t}$, the distribution of the disturbance will not differ between these observations, which suffices to generate the desired conditional moment inequalities. More formally consider the following alternative to Assumption 1.

Assumption H *For any fixed v in the support of $v_{i,s}$ and $v_{i,t}$,*

$$\varepsilon_{i,s}|x_{i,s}, x_{i,t}, \lambda_i, v_{i,s} = v_{i,t} = v \ \sim \ \varepsilon_{i,t}|x_{i,s}, x_{i,t}, \lambda_i, v_{i,s} = v_{i,t} = v.$$

This assumption augments the conditioning set of Assumption 1 to also include a common value v for the two observations. Under Assumption H, the distribution of the disturbance vector is identical whenever $v_{i,s} = v_{i,t}$. As a result we can derive conditional probability inequalities between the two observations exactly as in (13) with the conditioning set altered to $\Omega_{s,t} = \{x_{i,s}, x_{i,t}, \lambda_i, v_{i,s} = v_{i,t} = v\}$. This leads to the following proposition.

Proposition 2′′ *For any individual i making choices by maximizing (2), if Assumption H is satisfied then, for $s \neq t$ and any fixed v in the support of $v_{i,s}$ and $v_{i,t}$,*

$$0 \ \leq \ E\Big[m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta_0) \mid x_{i,s}, x_{i,t}, v_{i,s} = v_{i,t} = v\Big]$$

for $w = 0, 1, \ldots, K(x_{i,s}, x_{i,t}, \theta_0) - 2$, a.s. $(x_{i,s}, x_{i,t}, v_{i,s} = v_{i,t} = v)$.

Proposition 2′′ was motivated with a conditional heteroscedasticity example. When $v_{i,t}$ is correlated with $x_{i,t}$, conditional heteroskedasticity is just a particular form of dependence between regressors and disturbances. Other cases of dependence between regressors and disturbances can be handled by Proposition 2′′. Suppose there is a concern about "endogeneity" creating a violation of Assumption 1, e.g. some correlation between one or more disturbances and the regressors creating a failure of group homogeneity. Provided that a "control variable" v is available satisfying Assumption H, then Proposition 2′′ can be used to partially identify the model's parameters. The control variable must have the property that once we condition on it, the entire distribution of $\varepsilon_{i,t}$ conditional on $(x_{i,t}, x_{i,s}, v_{i,s} = v_{i,t} = v)$ is identical to that of $\epsilon_{i,s}$. However given a control variable with this property, conditioning on it will control for the endogeneity and allow for meaningful comparisons of corresponding choice probabilities. This illustrates the value of having variables beyond the covariates of the index function included in the conditioning set for the group homogeneity assumption.

A few further points about this proposition are worth noting. The choice equivalence sets in this proposition are the same as in Proposition 2, and hence determined only by $(x_{i,s}, x_{i,t})$ (and not by the value of $v_{i,s}$ and $v_{i,t}$). Still, the conditioning set includes the fixed value of v and so instrument functions will generally depend on v. Also the conclusion in this proposition depends on finding a pair of agents (i, s) and (i, t) with the same fixed values of $v_{i,s}$ and $v_{i,t}$. If the $v_{i,t}$ distribution is discrete then one could implement such a condition directly. If $v_{i,t}$ is continuously distributed then some smoothness in the conditional expectations would generally be needed to make use of this conditional moment inequality for inference or estimation. Finally, though in the case of conditional heteroskedasticity, the control variables $v_{i,t}$ might be observed directly, in other cases the control variables are likely to be determined in a first stage (typically using some instruments). If the control variable had to be estimated in a first stage, then one might further need to consider a version of the generated regressor approach mentioned in the next section.

4.2 Set-Valued Regressors

We consider the situation where one or more of the regressors that enter into random utility are not directly observed by the econometrician. Instead, the econometrician observes a set or region that is known to include the regressor value. Two familiar examples are cases where the regressor is: (i) income (or wealth) and all the econometrician knows is that the income of each observation lies in particular intervals; and (ii) the distance from home to a service (or retail) outlet when the home location is only observed as a zip code (with known geographic boundaries). As noted below an analogous construction to the one we provide here can be used for inference when there is a set that is known to contain the true value of the regressors with arbitrarily large probability (as is often the case when there are "generated regressors", or regressors whose values depend upon an estimated parameter).

There is a substantial applied and econometric literature dealing with interval-valued regressors. Manski and Tamer (2002) review some of that literature and provide bounds on a regression function under a monotonicity assumption with respect to the interval-observed variable. We consider multinomial choice (so the set-valued regressors could contribute to several index functions). We also allow for more than one covariate to be observed by interval or region, and do not impose monotonicity of the index functions. Of course, additional structure of some form (e.g. montonicity, or knowledge of the underlying distribution of the set-valued regressor, see Pollmann 2014) might provide more identifying power than our less restrictive framework.

As before, we simplify the notation by subsuming the index i for this discussion. For a given agent t, suppose that instead of observing the covariate x_t, our observables (say x_t^o) only tell us that the covariate x_t is contained in a set $\mathcal{X}_t$ with probability one. In our examples x_t^o would contain the endpoints of the intervals containing the true value of income, or the zip code of the agent's home. We will take that approach here and assume there is a mapping $\mathcal{X}$ from the observed random variables to the set that contains the true covariate, $\mathcal{X}_t = \mathcal{X}(x_t^o)$.[13] Some dimensions of $\mathcal{X}_t$ can be singletons (the dimensions of x_t that are observed without error) and some dimensions will be sets (intervals in the income example, but more complicated sets in the zip code example). Strictly speaking, $\mathcal{X}_t$ need not even take the form of a Cartesian product of sets and singletons.

Since we have $\mathcal{X}_s$ and $\mathcal{X}_t$ for observations s and t, once we fix θ we can compare the difference of index functions for any pair of choices by considering all the possible values of the covariates in $\mathcal{X}_s$ and $\mathcal{X}_t$. Specifically, for a pair of choices d and c, we can check if the smallest value of $g_d(z_s, \theta) - g_c(z_s, \theta)$ for $z_s \in \mathcal{X}_s$ is greater than the largest value of

[13] This formulation allows us to avoid introducing set-valued random variables. Molchanov (2005) provides a rigorous treatment of random sets that could alternatively be employed.

$g_d(z_t, \theta) - g_c(z_t, \theta)$ for $z_t \in \mathcal{X}_t$. If this statement about choices d and c is true, then it must hold that $g_d(x_s, \theta) - g_d(x_t, \theta) > g_c(x_s, \theta) - g_c(x_t, \theta)$, where x_s and x_t are the true values of the covariate for agents s and t.

In section 2, we used a pairwise comparison of index function differences to define equivalence sets of choices that could be ordered. This led to choice probability inequalities that were formed from unions of these equivalence sets. The probability inequalities (equation (13)) were of the form $\Pr(y_s \in D \,|\, \Omega_{s,t}) \geq \Pr(y_t \in D \,|\, \Omega_{s,t})$ when $g_d(x_s, \theta_0) - g_c(x_s, \theta_0) > g_d(x_t, \theta_0) - g_c(x_t, \theta_0)$ for all $d \in D$ and $c \notin D$. We now extend this approach to allow for set-valued regressors.

Start with the sets $\mathcal{X}_s$ and $\mathcal{X}_t$ for observations s and t. For a given value of the parameter θ, we want to find a set of choices D that insure that

$$g_d(x_s, \theta) - g_c(x_s, \theta) > g_d(x_t, \theta) - g_c(x_t, \theta) \tag{18}$$

for all $d \in D$ and $c \notin D$. This is equivalent to finding D such that

$$0 < \min_{\substack{d \in D,\\ c \notin D}} \Big(g_d(x_s, \theta) - g_c(x_s, \theta) - [g_d(x_t, \theta) - g_c(x_t, \theta)] \Big).$$

Since the true regressors are not observed, we can ensure that this condition holds by checking that it holds for every pair of possible regressor values in the sets $\mathcal{X}_s$ and $\mathcal{X}_t$. That is, suppose D is a set of choices satisfying

$$0 < \inf_{\substack{z_s \in \mathcal{X}_s,\\ z_t \in \mathcal{X}_t}} \min_{\substack{d \in D,\\ c \notin D}} \Big(g_d(z_s, \theta) - g_d(z_t, \theta) - [g_c(z_s, \theta) - g_c(z_t, \theta)] \Big). \tag{19}$$

This condition insures that for the true (x_s, x_t), the index function difference for choices in D is larger than the index function difference for choices in the complement of D. In particular, when (19) holds for $\theta = \theta_0$, then $g_d(x_s, \theta_0) - g_c(x_s, \theta_0) \; > \; g_d(x_t, \theta_0) - g_c(x_t, \theta_0)$ for $d \in D$, $c \notin D$, which is sufficient for generating a moment inequality for the partitioned choices.

Suppose there are $K(\mathcal{X}_s, \mathcal{X}_t, \theta)$ partitions of the choice set into two groups satisfying the condition in (19) for the set D above. Denote these partitions by $D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)$, for $k = 1, \ldots, K(\mathcal{X}_s, \mathcal{X}_t, \theta)$. So each $D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)$ partitions the choice set into two mutually exclusive and exhaustive sets of choices, $\{c \in D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)\}$, and $\{c \notin D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)\}$. Since these partitions are constructed in a manner that ensures that the index function difference in (18) holds at the true covariate value, these partitions are a subset of the partitions available when the true covariate values are observed (the case considered in section 2). As a result we could order these partitions so that $D^k(\mathcal{X}_s, \mathcal{X}_t, \theta) \subset D^{k+1}(\mathcal{X}_s, \mathcal{X}_t, \theta)$, just as we did above.

One way to construct these partitions for a given $(\mathcal{X}_s, \mathcal{X}_t, \theta)$ is to first look at each choice separately as a candidate for D and see if equation (19) is satisfied for any of them. Say (19) holds for some choice d_a. Next search for a partition consisting of two choices. When looking for a pair of choices that satisfy equation (19), one can restrict attention to just the pairs that include d_a as one of the elements. Regardless of whether a two-choice partition is found, one can next move to three-choice sets (that include d_a as an element). Similarly if we had not found a singleton partition, we would next search over all couples and continue from there.

A simple example might illustrate both the details involved in constructing the partitions, and the loss of information caused by not observing the value of the regressor. Consider a region which is divided into zip codes by passing vertical and horizontal lines through a map to form squares of equal size. Each axis is partitoned into intervals. Let h index the position of the interval on the east-west axis, and l indexes its position on the north-south axis. Distance is measured by the Euclidean distance between locations. Zip code A has location $(h = x, l = l_1)$ and zipcode B has $(h = x + 2, l = l_1)$. Relative to zip code A, observations in zip code B have shorter distances to travel to any outlet in zip codes indexed by $(x + \tau, l)$ for $\tau > 2$ and all l, and a longer distance to travel to outlets in zip codes with $\tau < 0$ and all l. However, we will not be able to order distances for outlets in zip codes $(h = x + 1, l)$ and any l. Of course, even if we cannot order distances, we may still be able to order differences in utilities from going to different outlets for a given θ because of the differences in non-distance features of the outlets. Whether we can or not will depend on the precise form of the utility function and the location of the outlet (which is generally known).

We now use the partitions of choices given above to define moment inequalities,

$$m_k^{\mathcal{X}}(y_s, y_t, \mathcal{X}_s, \mathcal{X}_t, \theta) = \mathbf{1}\{y_t \in D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)\} - \mathbf{1}\{y_s \in D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)\}.$$

To derive the expectation of this moment (or equivalently the corresponding difference of probabilities), we formally specify the relationship (i) between $\mathcal{X}$ and the true value of the covariate, and (ii) between $\mathcal{X}$ and the disturbances underlying the random utility for each choice. We will assume that the disturbance distribution is conditionally independent of the observable variables given the (possibly unobserved) true covariates and fixed effects (see Manski and Tamer (2002) for a similar conditional mean version of this assumption).

Assumption 2 *Assume that for any* (s, t),

a) $$(x_{i,s}, x_{i,t}) \in \Big(\mathcal{X}(x_{i,s}^o), \mathcal{X}(x_{i,t}^o)\Big)$$

with probability one; and

b) $$\varepsilon_{i,t}\big|x_{i,s}, x_{i,t}, \lambda_i, x_{i,s}^o, x_{i,t}^o \;\sim\; \varepsilon_{i,t}\big|x_{i,s}, x_{i,t}, \lambda_i.$$

Assumption 2 can be used to derive conditional moment inequalities for the set-valued covariate case analogous to the inequalities given in Proposition 2. These conditional moment inequalities can be derived by the argument provided in (13). The equality in that argument that is said to follow "by Assumption 1" will now follow by Assumptions 1 and 2. This argument leads to the following proposition.

Proposition 3 *For any group of observations (indexed by i) who make choices by maximizing (2), if Assumptions 1 and 2 hold, then*

$$0 \ \leq \ E\Big[m_k^{\mathcal{X}}(y_{i,s}, y_t, \mathcal{X}(x_{i,s}^o), \mathcal{X}(x_{i,t}^o), \theta_0) \ \big| x_{i,s}^o, x_{i,t}^o\Big]$$

for $k = 1, \ldots, K(\mathcal{X}_{i,s}, \mathcal{X}_{i,t}, \theta_0)$*, a.s.* $(x_{i,s}^o, x_{i,t}^o)$.

The implications of Proposition 3 are similar to those discussed after Proposition 2. Also, note that since the inequalities in Proposition 3 are conditional on $(x_{i,t}^o, x_{i,s}^o)$, positive valued functions of these variables can be used as "instruments" to form unconditional moment conditions.

It is also worth noting that if part a) of Assumption 2 held uniformly over all observations with a fixed probability,[14] then the inequalities in Proposition 3 would hold with that same probability. This finding would enable inference for the generated regressor case when uniform confidence sets for the true covariate values can be constructed from the generated regressors.

5 Finite-Dimensional Disturbance Distributions

Proposition 2 provides identifying conditional moment inequalities for the *semiparametric* multinomial choice model with fixed effects. In particular, the focus in the above sections has been on the case where no parametric distributional assumptions are made on the disturbances. The conditional moment inequalities of Proposition 2 do not appear to readily allow for the incorporation of parametric distribution information on the disturbances. In this section, we consider the multinomial choice model with fixed effects as specified before in equations (2) to (4) when the researcher is willing to make parametric assumptions on the distribution of disturbances.

[14] That is, $\Pr\left(\cap_{i,s\neq t}\left\{(x_{i,s}, x_{i,t}) \in \left(\mathcal{X}(x_{i,s}^o), \mathcal{X}(x_{i,t}^o)\right)\right\}\right) \geq 1-\alpha$.

Our interest centers on the case where the number of observations per group is small relative to the number of groups. This is the case generating an incidental parameter problem when asymptotic approximations are considered. In the contrasting case where the number of observations per group is large, then the likelihood corresponding to the parametric disturbance distribution can be used directly for identification. This case appears frequently in the demand literature, see Berry, Levinsohn, and Pakes (1995) and the literature that followed.[15]

Under an i.i.d. logistic assumption on disturbances, Chamberlain (1980) develops a conditional likelihood method of identification and estimation. Though we will also assume a parametric distribution for the disturbance, we will not require that it be i.i.d. logistic. Moreover, we do not require that the parametric distributional assumption on disturbances satisfies the group homogeneity assumption, i.e. our Assumption 1, so that the conditional moment inequalities of Proposition 2 do not necessarily hold if the parametric distributional assumption violates group homogeneity. For example, our parametric distributional assumption could allow for and specify some form of within group heteroskedasticity.

Of course, if one specifies a parametric disturbance distribution that maintains the group homogeneity assumption, then the conditional moment inequalities of Proposition 2 could be used to augment the information available on θ_0 (or to form a specification test of the parametric assumption). We note that the parametric information is likely to be particularly useful when $\mathscr{D}$, or the cardinality of the choice set, is large. When $\mathscr{D}$ is large, the use of Assumption 1 and the inequalities in Proposition 2 alone might be expected to be relatively uninformative as the $\mathscr{D}$-dimensional disturbance distribution underlying the observed choices is left entirely free.

The following assumption of a parametric disturbance distribution replaces Assumption 1.

Assumption 3 *Given the conditioning set $(x_{i,s}, x_{i,t})$, for any $s \neq t$, the conditional joint distribution of $\varepsilon_{i,s}$ and $\varepsilon_{i,t}$ is given by $F(\varepsilon_{i,s}, \varepsilon_{i,t}|x_{i,s}, x_{i,t}, \gamma_0)$, where $\gamma_0 \in \mathbb{R}^K$.*

Assumption 3 will often be derived from a specification of the joint (conditional) distribution of $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,T_i})$. We note that any parametric distribution will do here. In particular, we do not require conditions that this specification satisfy any restrictions on either the joint distribution of the disturbances across choices in a given period, or on the joint distribution of the disturbance vector for given choices across periods.

Again we omit the index i for notational convenience. Consider the case where different choices are made by the two observations s and t; that is where $y_s = c$ and $y_t = d$ for $c \neq d$.

[15] In particular, see **?**), for the case where N is small relative to the size of the choice set, and Berry, Gandhi, and Haile (2013) for the case when some (but not many) couples of goods are complements.

If $y_s = c$ and $y_t = d$ then

$$g_c(x_s,\theta_0) - g_d(x_s,\theta_0) + \lambda_c - \lambda_d + \epsilon_{c,s} - \epsilon_{d,s} \geq 0 \geq g_c(x_t,\theta_0) - g_d(x_t,\theta_0) + \lambda_c - \lambda_d + \epsilon_{c,t} - \epsilon_{d,t}.$$

Rearranging allows us to eliminate the fixed effects and write

$$(\epsilon_{c,s} - \epsilon_{d,s}) - (\epsilon_{c,t} - \epsilon_{d,t}) \geq \Big(g_d(x_s,\theta_0) - g_d(x_t,\theta_0)\Big) - \Big(g_c(x_s,\theta_0) - g_c(x_t,\theta_0)\Big). \tag{20}$$

In contrast to Assumption 1, the conditioning set in Assumption 3, and hence the parametric distribution of disturbances specified there, does not depend on the fixed effects. As a result we can directly compute the probability of an (ϵ_s, ϵ_t) combination satisfying the inequality above for any given value of the parameter vector. Let that probability be denoted

$$p_{s,t}^{c,d}(x_s, x_t, \theta, \gamma) =$$
$$\int \mathbf{1}\Big\{(\epsilon_{c,s} - \epsilon_{d,s}) - (\epsilon_{c,t} - \epsilon_{d,t}) \geq \Big(g_d(x_s,\theta) - g_d(x_t,\theta)\Big) - \Big(g_c(x_s,\theta) - g_c(x_t,\theta)\Big)\Big\}\, dF(\varepsilon_s, \varepsilon_t | x_s, x_t, \gamma)$$

Since we have shown that the event in (20) holds whenever $y_s = c$ and $y_t = d$, when we evaluate $p_{s,t}^{c,d}(x_s, x_t, \theta, \gamma)$ at (θ_0, γ_0), this computed probability must be at least as large as the probability that the event $(y_s = c, y_t = d)$ occurs. So if, for choices $c \neq d$, we define,

$$m_F^{c,d}(y_s, y_t, x_s, x_t, \theta, \gamma) \equiv p_{s,t}^{c,d}(x_s, x_t, \theta, \gamma) - \mathbf{1}\{y_s = c, y_t = d\},$$

we have the following proposition.

Proposition 4 *For any set of observations $(i,t)_{t=1}^{T_i}$ making choices by maximizing (2), if Assumption 3 is satisfied then, for $s \neq t$,*

$$0 \;\leq\; E\Big[m_F^{c,d}(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta_0, \gamma_0) \;\Big|\; x_{i,s}, x_{i,t}\Big]$$

for all $c \neq d$, a.s. $(x_{i,s}, x_{i,t})$.

Proposition 4 yields $\mathscr{D}(\mathscr{D}-1)$ conditional moment inequalities for each (s,t). Each group (our i index) has $T_i(T_i-1)/2$ observation pair comparisons each of which can be used to form a sample analogue of these inequalities. The probability $p_{s,t}^{c,d}(x_s, x_t, \theta, \gamma)$ could possibly be computed directly based on the parametric distribution in Assumption 3, or approximated through simulation. The simulation approximation does not require simulation of orthant probabilities, it simply requires draws from the distribution of a pair of choice disturbances. Of course it is this difference between the pairwise disturbance probability $p_{s,t}^{c,d}(x_s, x_t, \theta_0, \gamma_0)$

and the corresponding orthant probability, or $\Pr(y_s = c, y_t = d|x_s, x_t)$, that accounts for the slackness in the conditional moment inequality in Proposition 4 and leads to partial identification of the parameters. On the other hand, computing $\Pr(y_s = c, y_t = d|x_s, x_t)$ from the model would seem to require a specification for the distribution of the group's choice specific fixed effects (λ_i) conditional on $(x_{i,s}, x_{i,t})$. Proposition 4 allows us to avoid such a specification, and simply approximate $\Pr(y_s = c, y_t = d|x_s, x_t)$ directly from the data.

As above, conditional moment inequality methods can be used for estimation and inference. When N is the appropriate limiting dimension and observations in different groups are independent, then $p^{c,d}_{s,t}(x_s, x_t, \theta, \gamma)$ can be approximated by a single simulation draw if the expectation in Proposition 4 is estimated by a sample average over groups. Finally, note that the extensions in sections 4.1 and 4.2 can be developed for the parametric case as well.

6 Conclusion

We have provided a new approach to identification for multinomial choice models. Our focus has been on models which allow for choice-specific fixed effects with a group (or panel) structure and a nonparametric distribution of disturbances only restricted to satisfy a group homogeneity assumption. The utility for each choice is assumed to be additively separable in (i) a function of the choice-specific fixed effect and a disturbance, and (ii) an index function of covariates and parameters. We show that this structure generates moment inequalities which can be used to generate set estimators for the parameter vector. The main modeling advantages of our approach are that parametric distributional assumptions on the disturbance across choices are not needed, and we allow choice-specific fixed effects that can differ arbitrarily across groups. The framework can also account for set-valued regressors, and certain forms of endogeneity.

The main disadvantage of our approach is that, in general, it only leads to partial identification. On the other hand, our semiparametric approach does not require estimation of orthant probabilities, and, as a result, is relatively easy to use. So, one might think of using it to generate relatively assumption free information on θ_0 which can be used to check whether any added structure is appropriate. Section 5 of the paper is helpful in this context. It considers the same random utility multinomial choice setting with fixed effects, but assumes the disturbance distribution is known up to a parameter vector which needs to be estimated. The parametric distribution need not satisfy the group homogeneity assumption we use in our semiparametric analysis but, if it does, the parametric model is nested in our semiparametric model. The parametric model generalizes the assumptions previously used in estimation of panel data models with group and choice-specific fixed effects (see Chamberlain 1980) by allowing the specification of parametric distribution to be unrestricted across choices and across time periods for a given choice. However, as in the semiparmetric case, the assumption of a parametric distribution for the disturbances will, in general, only generate partial identification of the parameters of interest.

Our results point to a number of avenues of future research. Throughout we have focused on obtaining information on θ_0. Typically the focus of empirical studies is not on θ_0 per se but rather on different implications of its value. For instance, in panel data discrete choice settings, one might be interested in conditional quantile or average structural effects, as in Chernozhukov, Fernández-Val, Hahn, and Newey (2013). In such cases, our conditional moment inequalities provide information on θ_0 that can be established in an initial step and used to improve the estimation of the various marginal effects by narrowing the range parameter values that have to be considered. In the analysis of demand systems one will often be interested in own and cross price elasticities, or the consumer surplus generated by changes in product characteristics (Berry, Levinsohn, and Pakes 2004). We leave the question of the extent to which our results can be incorporated into the analysis of these and other issues for future research.

References

Abrevaya, J. (1999): "Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, 93(2), 203–228.

——— (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics*, 95(1), 1–23.

Andrews, D., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Aradillas-López, A., A. Gandhi, and D. Quint (2013): "Testing Inequalities of Conditional Moments, with an Application to Ascending Auction Models," University of Wisconsin Working Paper.

Armstrong, T. (2011): "Asymptotically Exact Inference in Conditional Moment Inequality Models," Stanford University Working Paper.

Bar, H., and F. Molinari (2013): "Computational Methods for Partially Identified Models via Data Augmentation and Support Vector Machines," Cornell University Working Paper.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected Substitutes and Invertibility of Demand," *Econometrica*, 81(5), 2087–2111.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

——— (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112(1), 68–105.

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

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 47(1), 225–238.

——— (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18(1), 5–46.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., S. Lee, and A. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Chetverikov, D. (2011): "Adaptive Test of Conditional Moment Inequalities," MIT Working Paper.

Fox, J. (2007): "Semiparametric Estimation of Multinomial Discrete-Choice Models using a Subset of Choices," *RAND Journal of Economics*, 38(4), 1002–1019.

Han, A. (1987): "Nonparametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35(2-3), 303–316.

Handel, B. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 103(7), 2643–2682.

Heckman, J. (1981): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. Manski, and D. McFadden, pp. 114–178. MIT Press, Cambridge.

Ho, K., and A. Pakes (forthcoming): "Hospital Choice, Hospital Prices and Financial Incentives to Physicians," *American Economic Review*.

Honore, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60(3), 533–565.

Kiefer, N. (1998): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26(2), 646–679.

Kroft, K., F. Lange, and M. Notowidigdo (2013): "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 128(3), 1123–1167.

Lee, L.-f. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65(2), 381–428.

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

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), 357–362.

Manski, C., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, pp. 105–142. Academic Press, New York.

Molchanov, I. (2005): *Theory of Random Sets*. Springer, New York.

Pakes, A. (2014): "Behavioral and Descriptive Forms of Choice Models," *International Economic Review*, 55(3), 603–624.

Pakes, A., J. Porter, K. Ho, and J. Ishii (forthcoming): "Moment Inequalities and Their Application," *Econometrica*.

Pollmann, D. (2014): "Identification and Estimation with an Interval-censored Regressor when its Marginal Distribution is Known," Harvard University Working Paper.

Powell, J. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54(6), 1435–1460.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70(1), 147–165.

Yan, J. (2013): “A Smoothed Maximum Score Estimator for Multinomial Discrete Choice Models,” University of Wisconsin Working Paper.