Moment Inequalities for Semiparametric Multinomial Choice with Fixed Effects*

Ariel Pakes
Harvard University and NBER

Jack Porter
University of Wisconsin

June 27, 2013

Abstract

We propose a new approach to identification of multinomial choice models with a group (or panel) structure. The utility for each choice is additively separable in a choice-specific fixed effect, a disturbance, and an index function of covariates and parameters. Observations in the same group are assumed to share the same fixed effects. Special cases of our semiparametric model include Chamberlain’s (1980) conditional logit model for panel data problems with choice-specific fixed effects and models of product demand where markets are the grouping device, the within group observations are consumers, and the choice-specific fixed effects represent product level unobservables. We place no restriction on the variance-covariance of the disturbance vector across choices. The only restriction on the disturbances is a group homogeneity assumption. The main cost of the semiparametric flexibility is that the derived conditional moment inequalities only partially identify the index function parameters. The advantages of our framework are that (i) it is nonparametric in the joint distribution of

*Thanks...
the disturbance vector across choices, (ii) it overcomes the incidental parameter problem induced by choice-specific fixed effects whose cardinality can grow with sample size, and (iii) the model can be extended to allow for set-valued covariates and certain forms of endogeneity.
1 Introduction

This paper proposes a new approach to identification for semiparametric multinomial choice models with a group (or panel) structure. We take a standard random utility model of choice, where the utility for each choice is additively separable in a choice-specific fixed effect, a disturbance, and an index function of covariates and parameters. The data consists of individual-level observed choices and covariates, and has a group structure, where observations in the same group are assumed to share the same fixed effects.

This 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. 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.

We consider a semiparametric version of this choice model that does not place parametric distributional assumptions on the disturbances. The only restriction on the disturbances is a group homogeneity assumption. To understand how this assumption is used consider two individuals in the same group and the differences of their observable index functions for each choice. Rank these differences from largest to smallest by choice. We derive inequalities on the choice probabilities between the two individuals for subsets of the highest ranked choices. This derivation extends the idea for binary choice in Manski (1987) to the case with multiple indices and disturbances. 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 conditional moment conditions 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. Not surprisingly, the extent of this cost is bound to vary by application. On the other hand, there are several potential benefits to the semiparametric approach. Most obviously the index function differencing step eliminates the additive fixed effect (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. Also, this
framework places no restrictions on the joint distribution of disturbances across choices. In particular, the marginal distributions of the choice-specific disturbances can have bounded support or mass points, and can differ arbitrarily across choices. Perhaps more importantly, no restrictions are placed on 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. 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.

We also 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 covariates, and a related argument allows us to analyze models with generated regressors. A number of other extensions are also possible. We show that control variables can be incorporated into the framework to deal with some forms of dependence of the disturbance distribution on the covariates (i.e. we can account for certain types of conditional heteroskedasticity and/or regressor endogeneity). We also note that, with some additional assumptions, we could analyze models with errors in the right hand side variables. As in our main results, all of the extensions allow for choice specific fixed effects and a non-parameteric distribution of the disturbance parameter.

Multinomial discrete choice estimation has been applied across a number of areas including the analysis of differentiated product markets, college choice, and hospital choice. These applications have typically employed parametric approaches (McFadden 1974, 1984, Hajivassiliou and Ruud 1994, Keane and Moffitt 1998). Manski (1975) introduced a semi-parametric, maximum score approach to point identification and estimation for multinomial choice without choice-specific fixed effects. Assuming homogeneity of unobservable components across choices, Manski then uses differences of the observable, parametric component of random utility across choices for identification. Using Manski’s identification approach, Fox (2007) shows that exchangeability of the unobservable across choices is sufficient for identification, and Yan (2013) obtains the limiting distribution for a smoothed version of the multinomial maximum score estimator. Rather than imposing conditions on the joint distribution of the disturbances across choices our approach imposes homogeneity of that distribution across individuals in a group but leaves the distribution of disturbances across choices unrestricted. 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. The different identifying assumptions in each work are likely to be useful in different applications. Our framework is designed to deal
with choice specific fixed effects in group (or panel) data settings, while simultaneously al-
lowing complete flexibility in the choice disturbance distribution to avoid model-imposed
substitution patterns across choices, and generalizes to allow for the extensions noted above.

The paper is structured as follows. Section 2 begins with notation for the standard
random utility response model. We then present and discuss our assumptions on the distur-
bance vector, and conclude with a derivation of the moment inequalities implied by those
assumptions. In section 3, extensions of the basic framework to set-valued covariates and
cases of endogeneity are considered.

2 Conditional Moment Inequalities

2.1 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 \( i \). There
is an obvious analogy to traditional panel data where \((i, t)\) indexes individuals and time respec-
tively. However there are a number of other familiar applications. 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). 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 2012).

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 \( D \) and number these choices so that \( d = 1, \ldots, D \). We consider the
case of unordered response, where the numbering associated with each choice is arbitrary.\(^1\)
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.

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

\[
U_{d,i,t} = g_d(x_{i,t}, \theta_0) + \lambda_{d,i} + \varepsilon_{d,i,t} \quad (1)
\]

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 a choice-specific fixed effect which accounts for unobserved characteristics of choice \( d \)
\(^1\)Inequalities for models with ordered responses are considered in Pakes, Porter, Ho, and Ishii (2011).
that do not vary across $t$. The term $\varepsilon_{d,i,t}$ represents any remaining unobserved, idiosyncratic determinants of the random utility.

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

$$y_{i,t} \in \arg\max_d U_{d,i,t}.$$  \hfill (2)

where the argmax function generates the set of choices that maximizes random utility. If a single choice is the unique maximizer of random utility, then equation (2) 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. Necessarily, 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 that case the model in (2) 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 and generated 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 $(i,t)$. The index function $g_d(x_{i,t}, \theta_0)$ is general enough to allow for the usual linear multinomial logit functional form, $x_{i,t}'\theta_0$, 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 $\theta_0$.

We now introduce the key stochastic assumption that will lead to our identification conditions. Notationally we let $\varepsilon_{i,t} = (\varepsilon_{1,i,t}, \ldots, \varepsilon_{D,i,t})'$ and $\lambda_i = (\lambda_{1,i}, \ldots, \lambda_{D,i})'$.

**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:

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

Assumption 1 is a group homogeneity assumption on the disturbances. No parametric distributional restrictions are placed on the form of the (marginal or conditional) distribution of $\varepsilon_{i,t}$. Indeed, the distribution of these disturbances can have bounded or unbounded sup-
port 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). Finally we note that the familiar panel data model assumption of strict exogeneity, that is

\[
\varepsilon_{i,t} \mid x_{i,1}, \ldots, x_{i,T}, \lambda_i \sim \varepsilon_{i,t} \mid x_{i,1}, \ldots, x_{i,T}, \lambda_i,
\]

would also suffice in place of Assumption 1 for the results below. 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). Note that neither strict exogeneity, nor Assumption 1, restrict the covariances of the joint distribution of \((\varepsilon_{i,s}, \varepsilon_{i,t})\); the key requirement is that the marginal distributions of the two disturbance vectors be the same.

Assumption 1 is the same assumption on disturbances used in Manski (1987) for binary choice panel.\(^2\) It clearly does restrict dependence across \(t\) and the relationship between the disturbances and the covariates. Those restrictions are easiest to see when \((\varepsilon_{i,t}, x_{i,t})\) is independent across \(t\) (conditional on \(\lambda_i\)). Then Assumption 1 boils down to an assumption of independence between the disturbance \(\varepsilon_{i,t}\) and the covariate \(x_{i,t}\) conditional on \(\lambda_i\), and an assumption of identical distributions for \(\varepsilon_{i,t}\) across \(t\) conditional on \(\lambda_i\).\(^3\)

From this case, it is easy to see that conditional heteroskedasticity in the disturbances would constitute a violation of Assumption 1. One of our extensions (section 3.3) will allow some relaxation of this assumption when the heteroskedasticity takes on a known multiplicative form.

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

\(^2\)Strictly speaking, the disturbance in the Manski (1987) binary choice model would correspond to the difference of disturbances given here across choices.

\(^3\)Conditioning on \(\lambda_i\) insures that particular \(\lambda_i\) are not associated with differences across \(t\) in the distribution of the \(\{\varepsilon_{i,t}\}_t\) (conditional on \(x_{i,1}, \ldots, x_{i,T}\)).
utilities across \( t \) for that \( i \). That comparison does not depend on the joint distribution of disturbances across choices in any way. Moreover, though typical methods of estimation and inference will combine the information on \( \theta \) from different groups, the distribution of disturbances is allowed to vary in an arbitrary way across those groups.

### 2.2 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 \( \theta_0 \). 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 \( i \). 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

\[
\Pr(y_t = d | \Omega_t) = \Pr \left( \varepsilon_{d,t} \geq \max_{c \neq d} \left[ g_c(x_t, \theta_0) + \lambda_c - (g_d(x_t, \theta_0) + \lambda_d) \right] + \varepsilon_{c,t} \bigg| \Omega_t \right), \tag{3}
\]

where \( \Omega_t \) can be any conditioning set. This probability involves the difference of the index functions and the fixed effects 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 and choice effect differences inside the probability in (3) above. In particular, suppose that if we compare the expression inside the square brackets in (3) for observations \( s \) and \( t \), we find that for all \( c \neq d \)

\[
g_c(x_t, \theta_0) + \lambda_c - (g_d(x_t, \theta_0) + \lambda_d) > g_c(x_s, \theta_0) + \lambda_c - (g_d(x_s, \theta_0) + \lambda_d). \tag{4}
\]

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 (3) to match the conditioning set in Assumption 1,
we obtain
\[ \Pr(y_s = d|x_s, x_t, \lambda) \geq \Pr(y_t = d|x_s, x_t, \lambda). \]

In the model without choice fixed effects, the inequalities in (4) could be checked directly for a given value of \( \theta \). However, since the choice fixed effects are constant within a group, they difference out in within-group comparisons, and this enables us to check the inequality for two observations within the same group.

Note that equation (4) holds for all \( c \neq d \) if and only if
\[ g_d(x_s, \theta_0) - g_d(x_t, \theta_0) > g_c(x_s, \theta_0) - g_c(x_t, \theta_0). \] (5)

Formally
\[ d = \arg\max_c \left( g_c(x_s, \theta_0) - g_c(x_t, \theta_0) \right) \Rightarrow \Pr(y_s = d|x_s, x_t, \lambda) \geq \Pr(y_t = d|x_s, x_t, \lambda). \] (6)

This is the basic idea motivating our choice probability inequalities.

2.3 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 inequalities. 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, this in turn, 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 \( U_t = \{ U_{1,t}, \ldots, U_{D,t} \} \),

\[
\{ U_t : U_{d,t} > \max_{c \neq d} U_{c,t} \} \subseteq \{ U_t : y_t = d \} \subseteq \{ U_t : U_{d,t} \geq \max_{c \neq d} U_{c,t} \}.
\]  

(7)

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 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 it does the set relationships in equation (7) imply conditional moment inequalities that do not depend on the distribution for the disturbance vector.

In addition to random utility ties, there is another source of potential “ties”. Notice that in equations (4) - (6), 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 \( \theta \), one could consider a parameter value that equates index differences.

### 2.4 Implied Moment Inequalities

The probability inequality in (6) 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 agents \( s \) 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 \( d = 1, \ldots, D \). So, \( 1 \leq K(x_s, x_t, \theta) \leq D \), and, when we order the index function differences, there are \( K(x_s, x_t, \theta) \) distinct rank values.

Given a value of \( \theta \), let the choices corresponding to the minimum difference of index functions be denoted

\[
D^{(1)}(x_s, x_t, \theta) = \arg \min_{c \in \{1, \ldots, D\}} \left[ g_c(x_s, \theta) - g_c(x_t, \theta) \right].
\]

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

\[
D^{(K(x_s,x_t,\theta))}(x_s, x_t, \theta) = \arg \max_{c \in \{1, \ldots, D\}} \left[ g_c(x_s, \theta) - g_c(x_t, \theta) \right].
\]

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

\[
g_c(x_s, \theta) - g_c(x_t, \theta) < g_d(x_s, \theta) - g_d(x_t, \theta).
\]

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 ranks of these equivalence sets generate the desired results on relative conditional probabilities. For instance, we can directly extend the result in (6) to conclude that

\[
D^{(K_{s,t})}_{s,t} = \arg \max_{c \in \{1, \ldots, D\}} \left( g_c(x_s, \theta_0) - g_c(x_t, \theta_0) \right)
\]

implies

\[
\Pr \left( y_s = D^{(K_{s,t})}_{s,t} | x_s, x_t, \lambda \right) \geq \Pr \left( y_t = D^{(K_{s,t})}_{s,t} | x_s, x_t, \lambda \right),
\]

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 (9), we first extend the set relations given in (7) to subsets of
Let $D$ denote any non-empty set of choices. We will use the following relationships

$$
\bigcup_{d \in D} \left\{ U_t : U_{d,s} > \max_{c \not\in D} U_{c,s} \right\} \subseteq \{ U_t : y_t \in D \} \subseteq \bigcup_{d \in D} \left\{ U_t : U_{d,t} \geq \max_{c \not\in D} U_{c,t} \right\}.
$$

We now derive the probability inequality in (9) for the highest ranked equivalence set. Since the derivation will also suffice for additional cases to be introduced below, we let $D = D^{(K_{s,t})}$ 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

$$
\Pr (y_s \in D | \Omega_{s,t}) \geq \Pr \left( \bigcup_{d \in D} \left\{ U_s : U_{d,s} > \max_{c \not\in D} U_{c,s} \right\} \bigg| \Omega_{s,t} \right) \tag{(11)}
$$

$$
= \Pr \left( \bigcup_{d \in D} \left\{ \varepsilon_{d,s} > \max_{c \not\in D} \left[ g_c(x_s, \theta_0) + \lambda_c - (g_d(x_s, \theta_0) + \lambda_d) + \varepsilon_{c,s} \right] \right\} \bigg| \Omega_{s,t} \right)
$$

$$
\geq \Pr \left( \bigcup_{d \in D} \left\{ \varepsilon_{d,s} \geq \max_{c \not\in D} \left[ g_c(x_s, \theta_0) + \lambda_c - (g_d(x_s, \theta_0) + \lambda_d) + \varepsilon_{c,s} \right] \right\} \bigg| \Omega_{s,t} \right)
$$

$$
= \Pr \left( \bigcup_{d \in D} \left\{ \varepsilon_{d,t} \geq \max_{c \not\in D} \left[ g_c(x_t, \theta_0) + \lambda_c - (g_d(x_t, \theta_0) + \lambda_d) + \varepsilon_{c,t} \right] \right\} \bigg| \Omega_{s,t} \right)
$$

$$
= \Pr \left( \bigcup_{d \in D} \left\{ U_t : U_{d,t} \geq \max_{c \not\in D} U_{c,t} \right\} \bigg| \Omega_{s,t} \right)
$$

$$
\geq \Pr \left( y_t \in D | \Omega_{s,t} \right).
$$

The first and last inequalities follow by the set inclusions in (10). 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 $c \not\in D^{(K_{s,t})}$, equation (8) gives the strict inequality

$$
g_c(x_t, \theta_0) + \lambda_c - (g_d(x_t, \theta_0) + \lambda_d) > g_c(x_s, \theta_0) + \lambda_c - (g_d(x_s, \theta_0) + \lambda_d).
$$

To obtain the probability inequality in (11), 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

\[ D^{(K_s,t-w-1)} \cup \ldots \cup D^{(1)}, \text{ for } w = 0, \ldots, D - 2. \]

In fact, setting \( D = D^{(K_s,t)} \cup \ldots \cup D^{(K_s,t-w)} \), then the derivation in (11) yields

\[
\Pr \left( \gamma_s \in D^{(K_s,t)} \cup \ldots \cup D^{(K_s,t-w)} | \Omega_{s,t} \right) \geq \Pr \left( \gamma_t \in D^{(K_s,t)} \cup \ldots \cup D^{(K_s,t-w)} | \Omega_{s,t} \right)
\]

for any \( w = 0, \ldots, D - 2 \).

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

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

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

\[
0 \leq E \left[ m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta_0) \mid x_{i,s}, x_{i,t} \right]
\]

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 \( \lambda_i \) 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 the empirical analogues of the relative probabilities of certain choice sets evaluated at or near \( \theta = \theta_0 \). Note that these moment inequalities may be consistent with other values of \( \theta \) as well, as we discuss below.

For a pair of observations \((x_{i,s}, y_{i,s})\) and \((x_{i,t}, y_{i,t})\) to provide information about \( \theta_0 \) through the moment functions used in the proposition, it must be the case that the moment functions evaluated at this pair of realized observations, \( m_w(y_{i,s}, y_{i,t}, x_{i,s}, x_{i,t}, \theta) \), vary to some extent with \( \theta \). Notice that when \( x_{i,s} = x_{i,t} \), all choices are in the same equivalence class regardless of value of \( \theta \), and hence the moment functions are identically zero for all \( \theta \). 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 \( \theta \). In either of these cases, this pair
of observations would not provide information about the true value $\theta_0$ through any empirical analog of a moment inequality derived from Proposition 1.

Now consider the case where $y_{i,s} \neq y_{i,t}$ and $x_{i,s} \neq x_{i,t}$, and assume that as $\theta$ 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 $\theta$, 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 $\theta$ 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 $\theta$ provides evidence that is consistent with the moment inequalities in Proposition 1. 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 $\theta$ 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 evidence against that parameter value.

It might also be instructive to compare using the set of conditional moment inequalities in Proposition 1 to what would happen if we focused only on the single conditional moment inequality generated by the highest ranked index difference, that is if we sufficed with the inequality in (6) that we started with in section 2.2. This would be equivalent to using only the first conditional moment, $m_0$, in Proposition 1. Now consider the information about $\theta_0$ 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 $\theta$ (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 $\theta$). When there are a large number of choices, a non-zero conditional moment function $m_0$ is unlikely to occur, regardless of the $\theta$ value. By using the conditional moment inequalities for the full range of choice rankings, we are able to glean considerably more information about $\theta$ 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 an “incidental parameter” problem. This problem is circumvented by considering the identifying implications of differences in the index functions. 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 estimation. An extreme case of \(T_i\) large and \(n\) small occurs when the data is cross sectional. Then \(n = 1\) and the \(\lambda\) 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.

Identified Set. If \(\Theta_{0,n}\) is defined as the set of parameters that satisfy the conditional moment inequalities in Proposition 1,

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

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

We have stated Proposition 1 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.\(^4\) 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

\(^4\)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 a 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

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

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, \(\Theta_{0,n} \subseteq \bigcap_{i=1}^{n} \bigcap_{s < t}^{T_i} \Theta_{0,i,s,t}\).
identified set will be a subset of $\Theta_{0,n}$ and yet will still contain $\theta_0$.

A sufficient condition for a value of $\theta$ to be in $\Theta_{0,n}$ is that, with probability one, $\theta$ and $\theta_0$ generate exactly the same ranking of index function differences.\(^5\) Though this condition is sufficient to insure $\theta \in \Theta_{0,n}$, it is not necessary. That is, there can be a $\theta$ which does not preserve the same index function difference rankings as $\theta_0$ 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 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 $\theta^*$. On the other hand, if the probability of $d_b$ for observation $(i, t)$ is larger than the probability of $d_b$ for observation $(i, s)$, then $\theta^* \notin \Theta_{0,n}$.

**Proposition 1 and Manski (1987).** Proposition 1 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 1 to the linear binary choice case (so $D = 2$ and, less important to the argument, $g_d(x, \theta) = x_d(\theta)$), which was the case considered in Manski (1987). Then Proposition 1 becomes

$$
(x_{2,i,s} - x_{2,i,t})'\theta_0 > (x_{1,i,s} - x_{1,i,t})'\theta_0 \implies \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) \geq \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t}). \tag{12}
$$

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

$$
(x_{2,i,s} - x_{2,i,t})'\theta_0 > (x_{1,i,s} - x_{1,i,t})'\theta_0 \iff \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) > \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t})
$$

and

$$
(x_{2,i,s} - x_{2,i,t})'\theta_0 = (x_{1,i,s} - x_{1,i,t})'\theta_0 \iff \Pr(y_{i,s} = 2|x_{i,s}, x_{i,t}) = \Pr(y_{i,t} = 2|x_{i,s}, x_{i,t}). \tag{13}
$$

Recall that the general result from the current paper (equation (12)) provides a **sufficient** condition for the difference in probabilities to be non-negative. The result in equation (13) 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

\(^5\)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 $\theta$ values will, at best, be up to an affine transform of the index functions (as is true in linear, parametric discrete choice models).
argument for point identification. In contrast, we do not need the unbounded support set assumption on the disturbance and only have sufficient conditions (and partial identification). The other 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.

Now consider the multinomial choice problem, and tentatively consider the additional assumption that each choice-specific disturbance has full support. As long as $D > 2$ our comparison of multiple index function differences will only suffice for the relative probability conclusions in our Proposition 1, it will not be necessary. For example consider comparing the choice of “a”. Clearly $Pr(y_s = a|x_s, x_t)$ could be higher than $Pr(y_t = a|x_s, x_t)$ without $a$ being the first highest ranked differences. It follows that our rank ordering conditions cannot be necessary conditions for our moment inequalities. So, in the multinomial choice case, a full support assumption does not provide a further refinement of the moment inequality results. Hence, we do not include this restriction.

*********I'm still thinking through the “refinement” statement...

Jack I don’t understand what refinement means.

**Computation and Estimation.** A computational advantage of basing estimation on moments derived from Proposition 1 is that it would not require the estimation of choice probabilities at different parameter values (this requires estimation of probabilities of complex regions for the value of the disturbance vector conditional on $\theta$). Estimation of these choice probabilities is standard in discrete choice estimation algorithms that specify a parametric distribution for the disturbance distribution and base the objective function 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 $\theta$ and this may impose an additional computational burden on the search
algorithm (a similar problem to that which occurs in parametric discrete choice problems which use frequency simulators). We leave open the question of whether one could do better with smoothed rankings, as in Yan (2013).

Since the empirical analog of the moments are straightforward to calculate, the main computational cost of our approach will be in employing conditional moment inequality methods for finding the estimator of the identified set and characterizing the aspects of its distribution needed for inference. For a discussion of these issues, see Bar and Molinari (2013).

The literature on inference on \( \theta_0 \) 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 is just a matter of 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 use based on equation (8). In particular, the difference of the index functions are already ranked for each \( \theta \).

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).

3 Extensions

We now extend the framework for developing conditional moment inequalities for partial identification that was presented in section 2. The extensions we consider allow for: (i) set-valued regressors, and (ii) some forms of both multiplicative heteroskedasticity and endogeneity.

3.1 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

\[ m_w(y_s, y_t, x_s, x_t, \theta) = g_{K,\varepsilon}(\theta) - w(x_s, \theta) - w(x_t, \theta) - [g_{1,\varepsilon}(x_s, \theta) - g_{1,\varepsilon}(x_t, \theta)]. \]

Footnote:

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

\[ m_w(y_s, y_t, x_s, x_t, \theta) = g_{K,\varepsilon}(\theta) - w(x_s, \theta) - w(x_t, \theta) - [g_{1,\varepsilon}(x_s, \theta) - g_{1,\varepsilon}(x_t, \theta)]. \]
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. monotonicity, or knowledge of the underlying distribution of the set valued regressor) 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 $x^*_t$) only tell us that the covariate $x_t$ is contained in a set $\mathcal{X}_t$ with probability one. In our examples $x^*_t$ 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, $\mathcal{X}_t = \mathcal{X}(x^*_t)$.

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 $\theta$ 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 $g_d(z_t, \theta) - g_c(z_t, \theta)$ 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 equiv-

---

7This formulation allows us to avoid introducing set-valued random variables. Molchanov (2005) provides a rigorous treatment of random sets that could alternatively be employed.
alence 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 (11)) were of the form \( \Pr(y_s \in D \mid \Omega_{s,t}) \geq \Pr(y_t \in D \mid \Omega_{s,t}) \) 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 \( c \not\in D \). 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 \( \theta \), we want to find a set of choices \( D \) that insure that

\[
g_d(x_s, \theta) - g_c(x_s, \theta) > g_d(x_t, \theta) - g_c(x_t, \theta)
\]

for all \( d \in D \) and \( c \not\in D \). This is equivalent to finding \( D \) such that

\[
0 < \min_{\substack{d \in D, \\ c \not\in D}} \left( g_d(x_s, \theta) - g_c(x_s, \theta) - [g_d(x_t, \theta) - g_c(x_t, \theta)] \right).
\]

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

\[
0 < \inf_{z_s \in \mathcal{X}_s, \; d \in D, \; z_t \in \mathcal{X}_t} \min_{\substack{c \not\in D}} \left( g_d(z_s, \theta) - g_d(z_t, \theta) - [g_c(z_s, \theta) - g_c(z_t, \theta)] \right) \tag{15}
\]

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 (15) holds for \( \theta = \theta_0 \),

\[
g_d(x_s, \theta_0) + \lambda_d - (g_c(x_s, \theta_0) + \lambda_c) > g_d(x_t, \theta_0) + \lambda_d - (g_c(x_t, \theta_0) + \lambda_c)
\]

for \( d \in D, \; c \not\in D \), 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 (15) 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 \( \{c \not\in D^k(\mathcal{X}_s, \mathcal{X}_t, \theta)\} \). Since these partitions are constructed in a manner that ensures that the index function difference in (14) 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 \((X_s, X_t, \theta)\) is to first look at each choice separately as a candidate for \(D\) and see if equation (15) is satisfied for any of them. Say (15) 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 (15), 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 find 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 partitioned 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 \(\tau \geq 2\) 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 \((h = x + 1, l)\), or for observations in adjacent zip codes. 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 \(\theta\) 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,

\[
m^X_k(y_s, y_t, X_s, X_t, \theta) = 1\{y_t \in D^k(X_s, X_t, \theta)\} - 1\{y_s \in D^k(X_s, X_t, \theta)\}.
\]

To derive the expectation of this moment (or equivalently the corresponding difference of probabilities), we formally specify the relationship between \(X\) and (i) the true value of the covariate, and (ii) between \(X\) 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 \((s,t)\),

\(a)\)

\((x_{i,s}, x_{i,t}) \in \left(\mathcal{X}(x_{i,s}^o), \mathcal{X}(x_{i,t}^o)\right)\)

with probability one; and
b) \[ \varepsilon_{i,t} \big| x_{i,s}, x_{i,t}, \lambda_i, x_{i,s}^o, x_{i,t}^o \sim \varepsilon_{i,t} \big| x_{i,s}, x_{i,t}, \lambda_i. \]

Under Assumption 2, the step in derivation (11) that is said to follow “by Assumption 1” will now follow by Assumptions 1 and 2. That proof then gives us the following proposition.

**Proposition 2** For any individual \( i \) making choices by maximizing (1), if Assumptions 1 and 2 hold, then

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

for \( k = 1, \ldots, K \left( \mathcal{X}_{i,s}, \mathcal{X}_{i,t}, \theta_0 \right), \text{ a.s.} \) \( (x_{i,s}, x_{i,t}) \).

The implications of Proposition 2 are similar to those discussed after Proposition 1. Also, note that since the inequalities in Proposition 2 are conditional on \( (x_{i,s}, x_{i,t}) \), positive valued functions of these variables can be used as “instruments” to form unconditional moment conditions.

**To Come.** A treatment of generated regressors.

### 3.2 Covariate Dependent Disturbance Distributions.

This section introduces a modification of Assumption 1 to accommodate certain forms of dependence between the distribution of disturbances \( (\varepsilon_{i,t}) \) and the regressors. We illustrate by considering the case of conditional heteroskedasticity. Suppose there is conditional heteroskedasticity 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 \( \sigma_d \) is unknown. Then 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 them, and that is all that is needed to generate the desired conditional moment inequalities. More formally consider the following alternative to Assumption 1.

**Assumption 1’** For any fixed \( v \) in the support of \( v_{i,s} \) and \( v_{i,t} \),

\[
\varepsilon_{i,s} \big| x_{i,s}, x_{i,t}, \lambda_i, v_{i,s} = v_{i,t} = v \sim \varepsilon_{i,t} \big| x_{i,s}, x_{i,t}, \lambda_i, v_{i,s} = v_{i,t} = v.
\]
This assumption augments the conditioning set of Assumption 1 to also include a common value $v$ for the two observations. Under Assumption $1'$ 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 (11) 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 1’** For any individual $i$ making choices by maximizing (1), if Assumption 1’ is satisfied then, for $s \neq t$ and any fixed $v$ in the support of $v_{i,s}$ and $v_{i,t},$

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

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 1’ 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 1’. Suppose there is a concern about “endogeneity” creating a violation of Assumption 1, i.e. one or more of the disturbances not being independent of the regressors. Provided that a “control variable” $v$ is available satisfying Assumption 1’, then Proposition 1’ 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 $\epsilon_{i,t}$ conditional on $(x_{i,t}, x_{i,s}, v_{i,s} = v_{i,t} = v)$ is identical to that of $\epsilon_{i,s}$ (for a detailed discussion of the use of control variables in econometrics, see Imbens and Wooldridge, forthcoming). 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.

A few further points about this proposition are worth noting. The choice equivalence sets in this proposition are the same as in Proposition 1, 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 last section.
4 Additional Identification Implications

Thus far, we have focused on identification of the parameter $\theta_0$, which, in turn, describes the index function, $g(\cdot, \theta_0)$. Of course, the ultimate objects of interest in multinomial choice problems could move well beyond the parameter $\theta_0$. Using the choice model for demand estimation, one could be interested in the effects of mergers or the introduction of new goods. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) [hereafter CFHN] consider average and quantile effects in nonlinear panel models.

We begin by showing that our moment inequalities can be used to refine the CFHN average effect bounds for multinomial choice problems.

5 Conclusion

We have provided a new identification strategy for multinomial choice models with a group (or panel) structure in which the utility for each choice is additively separable in a choice-specific fixed effect, a disturbance, and an index function of covariates and parameters. The main advantages of our approach are that it is nonparametric in the joint distribution of the disturbance vector across choices and it allows for choice-specific fixed effects. 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 it is relatively easy to use, so one might think of using it to get information on the parameter value which does not require a parametric specification of the disturbance vector before imposing additional structure on the disturbance distribution.

There are a number of issues left open in this version of our paper, some of which we are presently working on and will include in revisions soon. These issues include allowing for generated regressors and investigations of the extent to which our results can be used for prediction exercises and to identify the joint distribution of the disturbance vector. Moreover we have not presently considered how additional assumptions might be used in conjunction with our approach to enable an analysis of other familiar problems in discrete choice estimation; for example errors in right hand side variables. These are all topics which we hope will be addressed in future research.

References


