Moment Inequalities for Multinomial Choice with Fixed Effects

Ariel Pakes
Harvard University and NBER

Jack Porter
University of Wisconsin

August 12, 2019

Abstract

This paper proposes a new approach to identification of the semiparametric multinomial choice model with fixed effects. The framework employed is the semiparametric version of the traditional multinomial logit with fixed effects model (Chamberlain 1980). This semiparametric multinomial choice model places no restrictions on either the joint distribution of the random utility disturbances across choices or their within group (or across time) correlations. We show that a novel within-group comparison leads to a set of conditional moment inequalities which are the basis for two main identification results. First, we show that the derived conditional moment inequalities yield the sharp identified set for the random utility covariate coefficients, while avoiding the incidental parameter problem. Second, we add conditions on the covariates that are the multinomial analog of Manski's (1987) assumptions for the panel binary choice model. We show that these conditions, including an unbounded support for part of the covariate space, are sufficient to yield point identification. The key distinction from the binary choice framework is the multiple covariate index functions that determine the multinomial choice outcome. This distinction necessitates a new argument for point identification. The point identification result shows that, in fact, only a subset of the complete set of moment inequalities are needed for identification. Taken together, these results provide a thorough characterization of identification in the semiparametric multinomial choice model with fixed effects.

1 Introduction

This paper characterizes identification of the semiparametric multinomial choice model with fixed effects and a group (or panel) structure. A standard multinomial framework (McFadden 1974) is employed with random utility that is additively separable between unobservables, which include a disturbance and choice-specific fixed effects, and a linear covariate index. The key semiparametric assumption, replacing the multinomial logit specification (Chamberlain 1980), is a familiar group stationary condition on the disturbances. This assumption places no restrictions on either the joint distribution of the disturbances across choices or the correlation of disturbances across time. Under this specification, a novel within-group comparison leads to a set of conditional moment inequalities, which are the basis for our results on both point identification and sharp partial identification.

Our first findings address sharp identification in the semiparametric multinomial choice model with fixed effects. Under the group stationarity assumption alone, we find that our full set of derived conditional moment inequalities contains all of the model's potential identifying information in the sense that the bounds provided by these inequalities are sharp. Sharpness is shown by a constructive proof. In particular, given a distribution of observables and a parameter value in the identified set, we demonstrate that there exists a distribution of unobservables that can be combined with the parameter value to generate the given distribution of observables.

Next we turn to point identification. We propose a natural extension of Manski (1987)'s binary choice covariate assumptions to the multinomial setting. It is then shown that these assumptions are, in fact, sufficient for point identification of the coefficient parameter in the covariate index. Notably, this result holds with only two time periods so that the incidental parameter problem is solved completely by the use of within-variation. In fact, we find an even stronger result that the full collection of conditional moment inequalities is not needed for point identification. The conditional moment inequalities that are developed in this work are defined by various partitions of the set of choices. So, the full collection of conditional moment inequalities can be divided into subsets that correspond to partitions of the choices into two sets of fixed size. We show that each such subset of the conditional moment inequalities corresponding to a fixed size partitioning of the choices suffices for point identification. Depending on the number of choices being considered, this finding means that there is a particularly strong form of 'overidentification' in this model; there can be numerous non-overlapping sets of conditional moment inequalities that would each suffice for identification.

As these identification results are developed, we find substantive distinctions between the binary choice case and the general multinomial choice case considered here. Identification based on the maximum score criterion illustrates these differences. We first establish that the maximum score criterion can be derived as an aggregation of the conditional moments that make up our inequalities. In the binary choice model, we prove that the identified set determined by our conditional moment inequalities is exactly the set of parameters that maximize the maximum score criterion function. This new result shows that the maximum score criterion can be used for (sharp) identification (and hence its sample counterpart can be used for estimation) even when point identification fails in the binary choice panel model. While the maximum score criterion is robustly useful for identification in the binary choice case, we find that the opposite holds in multinomial choice with more than two choices. In the multinomial choice model with more than two choices, the defined extension of the maximum score criterion will not, in general, even be maximized at the true parameter value. This finding holds regardless of the validity or failure of point identification.

We are able to connect the distinction between findings for binary and more than two choices to differences in the implications of the conditional moment inequalities for the covariate indices in these two situations. In the binary choice model, a covariate index inequality is both necessary and sufficient for a corresponding conditional moment inequality. In the general multinomial case, the differenced covariate index rankings are only sufficient (and *not* necessary) to imply the corresponding conditional moment inequalities. This distinction between the binary choice model and the multinomial choice model leads to the differences in the ability of the maximum score criterion to identify the parameters. These distinctions also necessitate a new point identification argument for the general multinomial choice model.

Multinomial discrete choice models are extensively used in almost all fields that empirically analyze the determinants of agents' choices. Applications have typically employed parametric forms of the multinomial model. 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). 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. In a semiparametric multinomial setup, Shi, Shum, and Song (2018) focus on point identification and estimation using cyclic monotonicity. Kahn, Ouyang, and Tamer (2019) develop an approach to identification that can be used in both static and dynamic semiparametric multinomial choice models. Using a nonparametric multinomial choice model with endogeneity for the California Health Insurance Exchange, Tebaldi, Torgovitsky, and Yang (2018) identify and estimate bounds on counterfactuals. Gao and Li (2018) develops estimation and identification in a non-separable version of the panel multinomial choice model without restricting the joint distribution of disturbances.

This work also 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). Ahn, Ichimura, Powell, and Ruud (2017) develop an approach to identification and estimation of semiparametric index models that can also be used in various fixed effects cases. 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.

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 or time stationary assumption. Since the joint distribution of disturbances across choices is left unrestricted, the model contains no vestiges of independence of irrelevant alternatives or limits on cross price elasticities. Within group or across time disturbance correlation is also left completely unrestricted in this specification. The panel aspect of the model allows for an additive choice-specific fixed effect in the random utility specification. The fixed effects are allowed to be arbitrarily correlated with the observed covariates. We focus on the case with only two time periods (or group observations). The derived conditional moment inequalities are based on only within variation, so that the incidental parameter problem is fully circumvented in our identification results.

The paper is structured as follows. Section 2 sets up a semiparametric version the standard random utility model for multinomial choice with fixed effects. We then introduce our main stochastic disturbance assumption and derive a set of conditional moment inequalities. In section 3, we show that the conditional moment inequalities provide sharp bounds on the parameters of interests. In section 4, we find that when the conditional moment inequalities are combined with certain assumptions on covariates, point identification of the coefficient parameters can be achieved. Section 5 presents new results on the maximum score criterion and finds distinct behavior in the binary choice case and the multinomial choice case with at least three choices. Section 6 concludes. All proofs are in the Appendix.

2 Conditional Moment Inequalities for Multinomial Choice

2.1 Setup

The data will be assumed to have a group/panel structure, where $i = 1, \ldots, n$ indexes the groups and $t = 1, \ldots, T$ 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 (2014)). 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. Suppose that $d \in \{0, \ldots, \mathscr{D}\}$, so that the number of choices is $\mathscr{D} + 1$. We consider the case of unordered response, where the numbering associated with each choice is arbitrary,[1] and $2 \leq \mathscr{D} + 1 < \infty$.

Given covariates $x_{d,i,t}$ for each choice d associated with observation (i,t), the random utility for choices $d = 1, \ldots, \mathscr{D}$ takes the form

$$U_{d,i,t} = x'_{d,i,t}\theta_0 + \lambda_{d,i} + \varepsilon_{d,i,t}, \tag{1}$$

where the term $\lambda_{d,i}$ denotes choice-specific fixed effects which account for unobserved characteristics of choice d that do not vary across t. No restrictions are placed on the correlation between covariates $x_{d,i,t}$ and choice-specific fixed effects $\lambda_{d,i}$, so these fixed effect terms generate a potential incidental parameter problem. The term $\varepsilon_{d,i,t}$ represents any remaining unobserved, idiosyncratic determinants of the random utility. The covariates enter random utility through a linear index. The additive separability between the covariate index $x'_{d,i,t}\theta_0$ and the unobserved terms $\lambda_{d,i} + \varepsilon_{d,i,t}$ is critical to the results that follow. However, the additive separability between the fixed effect $\lambda_{d,i}$ and disturbance $\varepsilon_{d,i,t}$ could be relaxed. That is, $\lambda_{d,i} + \varepsilon_{d,i,t}$ could be replaced by a term of the form $f_d(\lambda_{d,i}, \varepsilon_{d,i,t})$, where f_d is an unknown nonlinear function for choice d.[2]

The random utility for choice $d = 0$ is normalized to be zero,

$$U_{0,i,t} = 0$$

so that the random utilities for all remaining choices are relative to choice 0. Similarly, we normalize $x_{0,i,t} = 0$ and $\lambda_{0,i} = \varepsilon_{0,i,t} = 0$.

The observed choice, $y_{i,t}$, for agent (i,t) maximizes the random utility $U_{d,i,t}$ over choices d. When a single choice uniquely maximizes random utility, then that choice is the observed choice for (i,t). We also allow for situations where there is a non-zero probability that two or more choices are maximize utility. This situation could occur if the distribution of $\varepsilon_{i,t}$ has mass points, as could be allowed under the flexible nonparametric assumptions on disturbances here, or when there are set-valued regressors.[3] To fully specify the choice decision, we adopt a simple rule for resolving ties among maximizing random utility choices.

[1]Inequalities for models with ordered responses are considered in Pakes, Porter, Ho, and Ishii (2015).

[2]Strictly speaking, under the assumptions below, the fixed effect could be absorbed into the disturbance without loss of generality.

[3]The set-valued regressor case is not covered in the present paper, but is analyzed in Pakes and Porter (2014). This situation is empirically relevant in problems where the researcher does not know the actual value of some of the regressors but does know that they lie in some set. Two familiar examples are when the regressor is: (i) income (or wealth) and all the econometrician knows is that the income of each observation lies in a particular interval; 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).

If choices d_1 and d_2 both maximize random utility ($U_{d_1,i,t} = U_{d_2,i,t} = \max_d U_{d,i,t}$) and $d_1 < d_2$, then assume that the larger of the choice numbers d_2 is the observed choice for (i,t). And, in general, if there are multiple utility maximizing choices, then the observed outcome is assumed to be the largest choice number among the utility maximizing choices.[4]

The setup thus far is a standard random utility formulation of multinomial choice except that, as in Chamberlain (1980), a choice-specific group fixed effect is included. It will be useful to establish notation for the mapping defined by this setup from the covariates, parameter θ_0, fixed effects, and disturbances to the observed outcome $y_{i,t}$. Let $x_{i,t} = (x'_{1,i,t}, \ldots, x'_{\mathscr{D},i,t})'$, $\lambda_i = (\lambda_{1,i}, \ldots, \lambda_{\mathscr{D},i})$, $\varepsilon_{i,t} = (\varepsilon_{1,i.t}, \ldots, \varepsilon_{\mathscr{D},i,t})$, and $U_{i,t} = (U_{1,i,t}, \ldots, U_{\mathscr{D},i,t})$. When it is helpful to be explicit about the dependence of random utility on its components, we will use the notation $U_d(x_{i,t}, \theta_0, \lambda_i, \varepsilon_{i,t})$ to denote $U_{d,i,t} = x'_{d,i,t}\theta_0 + \lambda_{d,i} + \varepsilon_{d,i,t}$. The observed outcome $y_{i,t}$ can also be written as a function of these same components: $y_{i,t} = y(x_{i,t}, \lambda_i, \varepsilon_{i,t}, \theta_0) = \max \operatorname{argmax}_d U_d(x_{i,t}, \theta_0, \lambda_i, \varepsilon_{i,t})$, where y is the mapping that represents the random utility formulation for multinomial choice given above.

Our identification results will correspond to the case where T is fixed, and, to further simplify the discussion, we will focus on the case where $T = 2$. We will denote the two time periods or observations within each group by s and t, rather than 1 and 2 to avoid confusion, especially in the variable subscripts, with the choices d which are numbered $0, \ldots, \mathscr{D}$.

The key stochastic assumption for this framework is within-group/time homogeneity of the disturbances. This assumption is sometimes called *strict exogeneity* and is a common condition imposed in static panel data models (Chernozhukov, Fernández-Val, Hahn, and Newey 2013).[5]

Assumption 1
(a) $(x_{i,s}, x_{i,t}, \lambda_i, \varepsilon_{i,s}, \varepsilon_{i,t})$ is independently and identically distributed for $i = 1, \ldots, n$;
(b) Given the conditioning set $(x_{i,s}, x_{i,t}, \lambda_i)$, the conditional distributions of $\varepsilon_{i,s}$ and $\varepsilon_{i,t}$ are the same:

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

The second part of the assumption mirrors the stochastic assumption made for panel data binary choice models in Manski (1987). No parametric distributional restrictions are placed on the distribution of $\varepsilon_{i,t}$. Note that $\varepsilon_{i,t}$ is, in general, a vector of individual choice disturbances, in contrast to the binary choice case. Importantly, for a given time t, the marginal

[4] Any non-stochastic rule for resolving utility maximizing ties will suffice for the results that follow.

[5] Mean independence and zero covariance forms of strict exogeneity also appear commonly in the literature, especially in linear panel data model cases. Here, the stronger conditional independence form of strict exogeneity will be employed.

distribution of these choice disturbances is allowed to vary arbitrarily across choices (d), and there are no restrictions on joint behavior of these 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. This assumption also allows the disturbances for the different choices to be freely correlated *across time*. Assumption 1 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).

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. Of course, the typical assumption of independence of disturbances and covariates across different s and t would suffice to satisfy Assumption 1.

Assumption 1(b) means that "within" variation could be useful for identification. 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 relative response probabilities by comparing the *observable* components of random utilities across t for that group i. This within-group comparison will not depend on the joint distribution of disturbances across choices in any way.

To simplify notation, below we eliminate the group i index with the understanding that all variables below are associated with the same group unless otherwise indicated.

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 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 a set of conditional moment inequalities.

Our moment inequalities are based on a within comparison of choice probabilities for individual/group i at times s and t. We can express the conditional probability of observing choice d at time t through the corresponding region of the disturbance space.

$$\mathcal{E}_{d,t} = \{\epsilon_t : y(x_t, \lambda, \epsilon_t, \theta_0) = d\} \tag{2}$$

And, given this definition, $\Pr(y_t = d | x_s, x_t, \lambda) = \Pr(\varepsilon_t \in \mathcal{E}_{d,t} \,|\, x_s, x_t, \lambda)$. To consider how variation in the covariates across time affects choice probabilities, it is useful to note the explicit dependence of the region $\mathcal{E}_{d,t}$ on the covariates, which follows from our rule for resolving utility maximizing ties:

$$\begin{aligned} \mathcal{E}_{d,t} \;=\; & \left\{ \epsilon_t : \epsilon_{d,t} \geq \max_{c<d} \left[\left(x'_{c,t}\theta_0 - x'_{d,t}\theta_0 \right) + (\lambda_c - \lambda_d) + \epsilon_{c,t} \right] \right\} \\ & \cap \left\{ \epsilon_t : \epsilon_{d,t} > \max_{c>d} \left[\left(x'_{c,t}\theta_0 - x'_{d,t}\theta_0 \right) + (\lambda_c - \lambda_d) + \epsilon_{c,t} \right] \right\}. \end{aligned} \tag{3}$$

We compare the time t regions $\mathcal{E}_{0,t}, \ldots, \mathcal{E}_{\mathscr{D},t}$ to the analogous regions at time s, $\mathcal{E}_{0,s}, \ldots, \mathcal{E}_{\mathscr{D},s}$. From this comparison, we will be able to show that for one of the $\mathscr{D}+1$ choices, the region at time s contains the corresponding region at time t. Moreover, the choice with this property is determined completely by the covariate indices. Assumption 1 then implies that the corresponding choice probability at time s will be at least as large as the choice probability at time t.

To find a choice with this special property, we order the covariate index differences across time by choice. In particular, find the choice with the largest change in covariate index:

$$d^* \;=\; argmax_c \;\; \left(x'_{c,s}\theta_0 \;-\; x'_{c,t}\theta_0 \right). \tag{4}$$

If there is more than one choice in the *argmax* set, then set d^* to any element of this set. Note that

$$\begin{aligned} & x'_{d^*,s}\theta_0 \;-\; x'_{d^*,t}\theta_0 \;\geq\; x'_{c,s}\theta_0 \;-\; x'_{c,t}\theta_0, \;\; \forall c \\ \implies \quad & x'_{c,t}\theta_0 \;-\; x'_{d^*,t}\theta_0 \;+\; (\lambda_c \;-\; \lambda_{d^*}) \geq\; x'_{c,s}\theta_0 \;-\; x'_{d^*,s}\theta_0 \;+\; (\lambda_c \;-\; \lambda_{d^*}), \;\; \forall c \end{aligned} \tag{5}$$

The latter covariate index differences on either side of the inequality are the same differences that define $\mathcal{E}_{d^*,t}$ and $\mathcal{E}_{d^*,s}$ in (3). And the inequality (5) ensures that

$$\mathcal{E}_{d^*,t} \subset \mathcal{E}_{d^*,s}$$

Hence,

$$\begin{aligned}
\Pr(y_s = d^* \,|\, x_s, x_t, \lambda) &= \Pr(\varepsilon_s \in \mathcal{E}_{d^*,s} \,|\, x_s, x_t, \lambda) \\
&= \Pr(\varepsilon_t \in \mathcal{E}_{d^*,s} \,|\, x_s, x_t, \lambda) \\
&\geq \Pr(\varepsilon_t \in \mathcal{E}_{d^*,t} \,|\, x_s, x_t, \lambda) \\
&= \Pr(y_t = d^* \,|\, x_s, x_t, \lambda). \qquad (6)
\end{aligned}$$

The first and last equalities follow from the definition of the disturbance regions in (2). The second equality follows from Assumption 1, and the inequality follows from the set inclusion derived above. Below we will extend the argument behind this inequality to generate additional conditional choice probability comparisons. Conditional probability inequalities like the one above can be translated into corresponding conditional moment inequalities, which, in turn, can be considered for identification of the parameter θ_0.

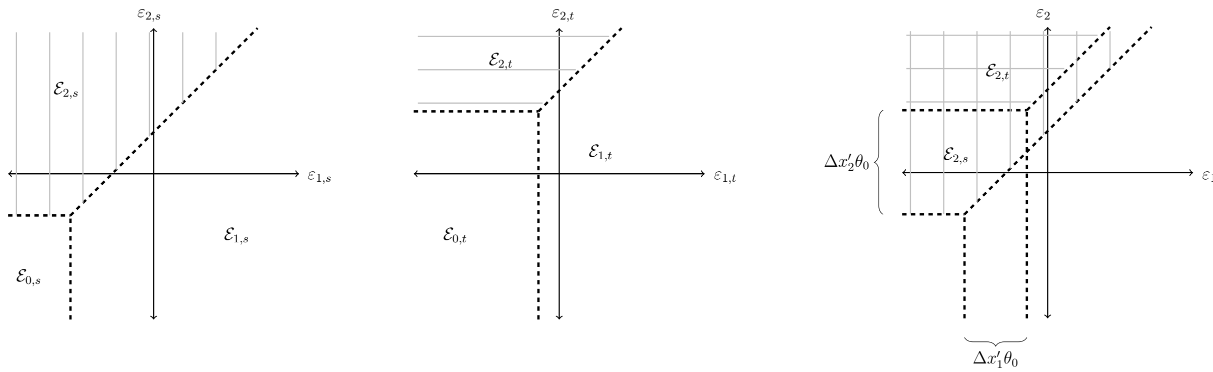
To illustrate the key intuition behind this inequality, consider the case with three choices and $d^* = 2$ implying that $\mathcal{E}_{2,t} \subset \mathcal{E}_{2,s}$. In Figure 1a, the time s disturbance space for ε_s is partitioned into regions $\mathcal{E}_{0,s}$, $\mathcal{E}_{1,s}$, and $\mathcal{E}_{2,s}$ corresponding to $y_s = 0, 1, 2$, respectively, and the vertical grey lines highlight region $\mathcal{E}_{2,s}$. A similar partitioning of the disturbance space at time t is shown in Figure 1b, and horizontal grey lines highlight region $\mathcal{E}_{2,t}$. Since ε_s and ε_t share the same (conditional) distribution and hence the same support set, these regions can be usefully superimposed in Figure 1c. From equation (3), the regions $\mathcal{E}_{0,t}$, $\mathcal{E}_{1,t}$, and $\mathcal{E}_{2,t}$ are a translation shift of the regions $\mathcal{E}_{0,s}$, $\mathcal{E}_{1,s}$, and $\mathcal{E}_{2,s}$ by $(x'_{1,s}\theta_0 - x'_{1,t}\theta_0, x'_{2,s}\theta_0 - x'_{2,t}\theta_0)$. The size of these shifts is apparent in Figure 1c, and, in the case illustrated, $x'_{2,s}\theta_0 - x'_{2,t}\theta_0 \geq x'_{1,s}\theta_0 - x'_{1,t}\theta_0 \geq 0$, it follows from (4) that $d^* = 2$. Starting from the nexus of the regions in Figure 1a, the direction of the translation shift is interior to $\mathcal{E}_{2,s}$, which ensures that $\mathcal{E}_{2,t} \subset \mathcal{E}_{2,s}$.

2.3 Implied Moment Inequalities

The probability inequality in (6) is based on the choice that maximizes the difference of covariate index functions. We can push this logic further to obtain similarly motivated inequalities based on a rank ordering of the covariate index function differences over the choices. For time periods s and t, start by ordering the difference of index functions by choice. This ordering can be used to partition the choices into a set of choices with larger index function differences and a set with smaller index function differences. For each such partition generated by index function differences based on the true parameter value θ_0, we will be able to generate corresponding choice probability inequalities.

Figure 1: Disturbance Regions for 3 Choices

$$\Delta x_2'\theta_0 \geq \Delta x_1'\theta_0 \geq 0 \quad \Longrightarrow \quad \mathscr{E}_{2,t} \subset \mathscr{E}_{2,s}$$



(a) (b) (c)

Given a value of θ and vectors of covariates x_s and x_t, we can partition the set of choices into two subsets corresponding to choices with larger and smaller index function differences. For instance, suppose D is a subset of choices, i.e. $D \subset \{0, 1, \ldots, \mathscr{D}\}$, and let D^c denote the remaining choices, $D^c = \{0, 1, \ldots, \mathscr{D}\}\backslash D$. If

$$\min_{d \in D} \Delta x_d'\,\theta \geq \max_{c \in D^c} \Delta x_c'\,\theta, \qquad \text{where } \Delta x_d = x_{d,s} - x_{d,t},$$

then D contains choices with larger index function differences and D^c contains choices with smaller index function differences. There are many possible partitions that could be formed in this way, and we collect the subsets corresponding to the larger index function differences as follows,

1 1 1

$$\overline{\mathbb{D}}(x_s, x_t, \theta) = \left\{ D \subset \{0, 1, \ldots, \mathscr{D}\} \,\middle|\, D, D^c \neq \varnothing,\ \min_{d \in D} \Delta x_d'\,\theta \geq \max_{c \in D^c} \Delta x_c'\,\theta \right\}. \tag{7}$$

Consider the case where no pair of choices share the same the index function difference, so that each choice has a distinct index function difference value. Then $\overline{\mathbb{D}}(x_s, x_t, \theta)$ will contain a set consisting of the choice corresponding to the largest index function difference ($\{d^*\}$ in the notation of the previous section). It will also contain a set consisting of the choices associated with the two largest index function differences, etc. So, in this case, $\overline{\mathbb{D}}(x_s, x_t, \theta)$ would contain exactly $\mathscr{D}$ sets with cardinalities 1, 2, …, $\mathscr{D}$. Moreover for any two sets $C, D \in \overline{\mathbb{D}}(x_s, x_t, \theta)$ where C has fewer elements than D, then $C \subset D$.

In general, of course it's also possible that index function differences for some choices will be equal. When this happens, $\overline{\mathbb{D}}(x_s, x_t, \theta)$ will contain more than $\mathscr{D}$ sets, and the sets in $\overline{\mathbb{D}}(x_s, x_t, \theta)$ will not be nested. As a simple example, suppose there are four choices, $\{0, 1, 2, 3\}$. And suppose the index function differences can be ordered as follows: $\Delta x_3' \theta > \Delta x_2' \theta = \Delta x_1' \theta > \Delta x_0' \theta$ (note that $\Delta x_0 = 0$). Then, $d^* = 3$ and $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{3\}, \{3, 2\}, \{3, 1\}, \{3, 2, 1\}\}$.

The next result shows that the argument used to obtain a probability choice inequality for d^* in the previous section can be extended to the choice sets contained in $\overline{\mathbb{D}}(x_s, x_t, \theta_0)$.

Proposition 1 *Suppose Assumption 1 holds. Then, for all* $D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0)$,

$$\Pr(y_s \in D \,|\, x_s, x_t, \lambda) \;\geq\; \Pr(y_t \in D \,|\, x_s, x_t, \lambda).$$

□

Proof: For any set of choices $C \subset \{0, 1, \ldots, \mathscr{D}\}$ and a choice d, define

$$\begin{aligned} \mathscr{E}_{d,C,t} \;&=\; \left\{ \epsilon_t : \epsilon_{d,t} \geq \max_{c \in C : c < d} \left[\left(x_{c,t}'\theta_0 - x_{d,t}'\theta_0 \right) \,+\, (\lambda_c - \lambda_d) + \epsilon_{c,t} \right] \right\} \\ &\quad \cap \left\{ \epsilon_t : \epsilon_{d,t} > \max_{c \in C : c > d} \left[\left(x_{c,t}'\theta_0 - x_{d,t}'\theta_0 \right) \,+\, (\lambda_c - \lambda_d) + \epsilon_{c,t} \right] \right\} \end{aligned}$$

The set $\mathscr{E}_{d,C,t}$ is the region of the time t disturbance space where choice d is preferred (in random utility terms) to all choices in C. Corresponding time s regions $\mathscr{E}_{d,C,s}$ are defined analogously.

Now take any $D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0)$. For any $d \in D$, $\Delta x_d' \theta_0 \geq \Delta x_c' \theta_0$ for all $c \in D^c$. Re-arranging, $x_{c,t}' \theta_0 - x_{d,t}' \theta_0 \;\geq\; x_{c,s}' \theta_0 - x_{d,s}' \theta_0$ for all $c \in D^c$. It follows that $\mathscr{E}_{d,D^c,t} \subset \mathscr{E}_{d,D^c,s}$ and this set inclusion is the main step to showing the desired probability inequality:

$$\begin{aligned} \Pr(y_s \in D \,|\, x_s, x_t, \lambda) \;&=\; \Pr\left(\varepsilon_s \in \bigcup_{d \in D} \mathscr{E}_{d,D^c,s} \,\middle|\, x_s, x_t, \lambda \right) \\ &=\; \Pr\left(\varepsilon_t \in \bigcup_{d \in D} \mathscr{E}_{d,D^c,s} \,\middle|\, x_s, x_t, \lambda \right) \\ &\geq\; \Pr\left(\varepsilon_t \in \bigcup_{d \in D} \mathscr{E}_{d,D^c,t} \,\middle|\, x_s, x_t, \lambda \right) \qquad (8) \\ &=\; \Pr(y_t \in D \,|\, x_s, x_t, \lambda) \end{aligned}$$

Holding (x_s, x_t, λ) fixed, $\{\varepsilon_s : y_s \in D\} = \cup_{d \in D} \mathscr{E}_{d,D^c,s}$ which is the argument behind the first equality. The last equality holds similarly. The second equality holds by Assumption 1. Finally, the inequality in (8) holds since the set inclusion $\mathcal{E}_{d,D^c,t} \subset \mathcal{E}_{d,D^c,s}$ for all $d \in D$ implies

$$\bigcup_{d\in D} \mathcal{E}_{d,D^c,t} \subset \bigcup_{d\in D} \mathcal{E}_{d,D^c,s}. \tag{9}$$

$\square$

The probability inequalties obtained in Proposition 1 can be straightforwardly translated into corresponding moment inequalities, as follows. For any $D \subset \{0, 1, \ldots, \mathscr{D}\}$, define $m_D(y_s, y_t) = \mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\}$. Then, Proposition 1 concludes that $E[m_D(y_s, y_t) \,|\, x_s, x_t, \lambda] \;\geq\; 0 \quad \forall D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0)$. Taking the expectation with respect to λ conditional on x_s, x_t yields conditional moment inequalities expressed in terms of the observables (y_s, y_t, x_s, x_t):

$$E[m_D(y_s, y_t) \,|\, x_s, x_t] \;\geq\; 0 \qquad \forall D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0). \tag{10}$$

This set of conditional moment inequalities will be the key to the identification arguments that follow.

3 Sharp Identification

The conditional moment inequalities in (10) generated by Proposition 1 depend only on the observable variables, (y_s, y_t, x_s, x_t) with distribution F_{y_s,y_t,x_s,x_t}. So, the corresponding identified set can be defined as follows:

$$\Theta_0 = \Theta_0(F_{y_s,y_t,x_s,x_t}) = \{\theta \in \Theta \,|\, E[m_D(y_s, y_t) \,|\, x_s, x_t] \;\geq\; 0 \;\; \forall D \in \overline{\mathbb{D}}(x_s, x_t, \theta) \; a.s. \; (x_s, x_t)\}.$$

Under Assumption 1 alone, Θ_0 will *not* generally be a point (up to scale). However, the below result shows that the conditional moment inequalities defining Θ_0 do, in fact, contain all available information about the parameter in the sense that the identified set, Θ_0, provides sharp bounds on the parameter.

Given random variables $(x_s, x_t, \lambda, \varepsilon_s, \varepsilon_t)$ satisfying Assumption 1 and a value of the parameter $\theta \in \Theta$, the multinomial choice framework defines outcomes $y_s = y(x_s, \lambda, \varepsilon_s, \theta)$ and $y_t = y(x_t, \lambda, \varepsilon_t, \theta)$. Then, the observable random variables from the multinomial choice model satisfying Assumption 1 are simply (y_s, y_t, x_s, x_t) with distribution F_{y_s,y_t,x_s,x_t}. We can collect all such values of the parameter and observable distributions:

$$\begin{aligned}\mathcal{M} \;=\; \{(\theta, F_{y_s,y_t,x_s,x_t}) \mid \;& (x_s, x_t, \lambda, \varepsilon_s, \varepsilon_t) \text{ satisfies Assumption } 1, \theta \in \Theta, \\ & y_s = y(x_s, \lambda, \varepsilon_s, \theta), y_t = y(x_t, \lambda, \varepsilon_t, \theta), (y_s, y_t, x_s, x_t) \sim F_{y_s,y_t,x_s,x_t}\}\end{aligned}$$

Let F_{y_s,y_t,x_s,x_t} be any observable distribution from the multinomial choice framework under Assumption 1, i.e. for some $\theta \in \Theta$, $(\theta, F_{y_s,y_t,x_s,x_t}) \in \mathcal{M}$. Then the sharp identified set is simply the projection of $\mathcal{M}$ onto Θ for the given observable distribution,

$$\Theta_S = \Theta_S(F_{y_s,y_t,x_s,x_t}) = \{\theta \in \Theta \mid (\theta, F_{y_s,y_t,x_s,x_t}) \in \mathcal{M}\}$$

Sharpness of the identified set Θ_0 is simply that $\Theta_0 = \Theta_S$. More formally, let

$$\mathcal{F}_{ob} = \{F_{y_s,y_t,x_s,x_t} \mid (\theta, F_{y_s,y_t,x_s,x_t}) \in \mathcal{M} \text{ for some } \theta \in \Theta\}.$$

So, $\mathcal{F}_{ob}$ is the set of all possible multinomial choice observable distributions generated by some distribution of unobservables satisfying Assumption 1 and some parameter value $\theta \in \Theta$. Then, we have the following result.

Theorem 2 *Under Assumption 1, Θ_0 is sharp. That is,*

$$\Theta_0(F_{y_s,y_t,x_s,x_t}) = \Theta_S(F_{y_s,y_t,x_s,x_t})$$

for all $F_{y_s,y_t,x_s,x_t} \in \mathcal{F}_{ob}$.

□

Theorem 2 is shown by a constructive proof, see Appendix for details. Fix any distribution of observables, $F_{y_s,y_t,x_s,x_t} \in \mathcal{F}_{ob}$. Let Θ_0 and Θ_S denote the identified sets associated with this observable distribution, $\Theta_0 = \Theta_0(F_{y_s,y_t,x_s,x_t})$ and $\Theta_S = \Theta_S(F_{y_s,y_t,x_s,x_t})$. It is straightforward to establish that $\Theta_S \subset \Theta_0$ (using Proposition 1). So, the main argument for Theorem 2 is to show the set inclusion in the other direction, $\Theta_0 \subset \Theta_S$.

Take any $\theta \in \Theta_0$. We exhibit a conditional distribution $(\lambda^*, \varepsilon_s^*, \varepsilon_t^*) \mid x_s, x_t$ such that $(x_s, x_t, \lambda^*, \varepsilon_s^*, \varepsilon_t^*)$ satisfies Assumption 1 and $(y_s^*, y_t^*, x_s, x_t) \sim F_{y_s,y_t,x_s,x_t}$ where $y_s^* = y(x_s, \lambda^*, \varepsilon_s^*, \theta)$ and $y_t^* = y(x_t, \lambda^*, \varepsilon_t^*, \theta)$. Then, $(\theta, F_{y_s,y_t,x_s,x_t}) = (\theta, F_{y_s^*,y_t^*,x_s,x_t}) \in \mathcal{M}$ and so $\theta \in \Theta_S$.

Let (x_s, x_t) be any pair of covariate values in the support of the joint distribution $\mathcal{X}_{s,t}$. We need to choose $(\lambda^*, \varepsilon_s^*, \varepsilon_t^*) \mid x_s, x_t$ such that $\Pr(y_s^* = d, y_t^* = d' \mid x_s, x_t) = \Pr(y_s = d, y_t = d' \mid x_s, x_t)$. Setting $\lambda^* = 0$, $\Pr(y_s^* = d, y_t^* = d' \mid x_s, x_t)$ is determined by the behavior of $(\varepsilon_s^*, \varepsilon_t^*) \mid x_s, x_t$ on certain "choice-determining" regions of $\mathbb{R}^{2\mathcal{D}}$. We further subdivide these regions so that symmetry can be imposed on the corresponding discrete marginal distributions to satisfy Assumption 1.

Let $R_{d,d'} = \{\varepsilon \,|\, y(x_s, \theta, \lambda^* = 0, \varepsilon) = d\} \cap \{\varepsilon \,|\, y(x_t, \theta, \lambda^* = 0, \varepsilon) = d'\}$. These sets can be used to subdivide the choice-determining sets on $\varepsilon_s^* \,|\, x_s, x_t$ and to similarly subdivide the choice-determining sets on $\varepsilon_t^* \,|\, x_s, x_t$. That is,

$$\{\varepsilon_s^* \,|\, y(x_s, \theta, \lambda^* = 0, \varepsilon_s^*) = d\} = \bigcup_{d'=0}^{\mathscr{D}} R_{d,d'} \quad \text{and} \quad \{\varepsilon_t^* \,|\, y(x_t, \theta, \lambda^* = 0, \varepsilon_t^*) = d'\} = \bigcup_{d=0}^{\mathscr{D}} R_{d,d'}.$$

So $\Pr(y_s^* = d \,|\, x_s, x_t) = \sum_{d'=0}^{\mathscr{D}} \Pr(\varepsilon_s^* \in R_{d,d'} \,|\, x_s, x_t)$ and similarly $\Pr(y_t^* = d' \,|\, x_s, x_t) = \sum_{d=0}^{\mathscr{D}} \Pr(\varepsilon_t^* \in R_{d,d'} \,|\, x_s, x_t)$. From these expressions, we see that the sets $R_{d,d'}$ can be used to describe the marginal behavior of y_s^* (and y_t^*), and in fact provide a device for imposing the homogeneity in ε_t^* across time t as required by Assumption 1(b).

Additionally, the Cartesian products of sets of the form $R_{d,d'}$ can be used to describe the joint behavior of y_s^* and y_t^*. So,

$$\Pr(y_s^* = d, y_t^* = d' \,|\, x_s, x_t) = \sum_{d''=0}^{\mathscr{D}} \sum_{d'''=0}^{\mathscr{D}} \Pr((\varepsilon_s^*, \varepsilon_t^*) \in R_{d,d''} \times R_{d''',d'} \,|\, x_s, x_t).$$

Now, let $q^*_{d,d' \times d'',d'''} = \Pr((\varepsilon_s^*, \varepsilon_t^*) \in R_{d,d'} \times R_{d'',d'''} \,|\, x_s, x_t)$. Then, we can translate our problem of exhibiting a conditional distribution $(\lambda^*, \varepsilon_s^*, \varepsilon_t^*) \,|\, x_s, x_t$ that both satisfies Assumption 1 and generates an observable distribution matching F_{y_s, y_t, x_s, x_t} into a problem of finding a solution to the system of linear equations below. Here, we treat $\Pr(y_s = d, y_t = d' \,|\, x_s, x_t)$ as known and seek a nonnegative solution for $q^*_{d,d'' \times d',d'''}$:

$$\Pr(y_s = d, y_t = d' \,|\, x_s, x_t) = \sum_{d''=0}^{\mathscr{D}} \sum_{d'''=0}^{\mathscr{D}} q^*_{d,d'' \times d''',d'} \tag{11}$$

$$\sum_{d''=0}^{\mathscr{D}} \sum_{d'''=0}^{\mathscr{D}} q^*_{d,d' \times d'',d'''} = \sum_{d''=0}^{\mathscr{D}} \sum_{d'''=0}^{\mathscr{D}} q^*_{d'',d''' \times d,d'} \tag{12}$$

If $q^*_{d,d' \times d'',d'''}$ satisfies (11), then $(y_s^*, y_t^*, x_s, x_t) \sim F_{y_s, y_t, x_s, x_t}$, as desired. If $q^*_{d,d' \times d'',d'''}$ satisfies (12), then the conditional disturbance distribution can be constructed to satisfy Assumption 1. A simple way to achieve this construction is to choose a point in each region, $r_{d,d'} \in R_{d,d'}$. Then choose the distribution of $(\varepsilon_s^*, \varepsilon_t^*) \,|\, x_s, x_t$ to be discrete with $\Pr((\varepsilon_s^*, \varepsilon_t^*) = (r_{d,d'}, r_{d'',d'''}) \,|\, x_s, x_t) = q^*_{d,d' \times d'',d'''}$.

The last step is to show existence of a nonnegative solution for $q^*_{d,d' \times d'',d'''}$ satisfying (11) and (12). Existence is established for an equivalent dual problem, using Farkas' Lemma, see

Appendix.

There are several interesting features of the constructed distribution that proves Theorem 2. First, there are additional constraints not apparent in equations (11) and (12). In particular, from the proof of Proposition 1, we know that some of the regions $R_{d,d'}$ will be empty, and so any corresponding $q^*_{d,d' \times d'',d'''}$ or $q^*_{d'',d''' \times d,d'}$ will be zero. For example, suppose d^* is the covariate index difference maximizing choice based on θ, as defined in section 2.2. Then, $\{\varepsilon^*_t \,|\, y(x_t, \theta, \lambda^* = 0, \varepsilon^*_t) = d^*\} \subset \{\varepsilon^*_s \,|\, y(x_s, \theta, \lambda^* = 0, \varepsilon^*_s) = d^*\}$. The choice-determining set for d^* at time t is contained in the choice-determining set for d^* at time s, which is the basic idea behind the first conditional probability inequality derived in (6). Fixing d^* to be defined as in section 2.2, it follows that for $d \neq d^*$, $R_{d,d^*} = \varnothing$. Additional constraints of this type have to be accounted for in the solutions to (11) and (12).

Second, the fixed effects are set to zero in the constructed distribution and essentially play no role. Fixed effect variation is not needed to match the simulated distribution to the given distribution of observables. This feature also reflects the fact that only within variation is used in the identification through the conditional moment inequalities.

Third, note that some across time correlation in the distribution of $(\varepsilon^*_s, \varepsilon^*_t) \,|\, x_s, x_t$ is, in general, needed to solve equations (11) and (12). Notice that satisfaction of Assumption 1 is achieved by matching the distributions of $\varepsilon^*_s \,|\, x_s, x_t$ and $\varepsilon^*_t \,|\, x_s, x_t$. However, the full flexibility from the conditional joint distribution $(\varepsilon^*_s, \varepsilon^*_t) \,|\, x_s, x_t$ is needed to match the simulated distribution to the given observable distribution, which can include correlation between choices across time. In particular, if we impose conditional independence between ε^*_s and ε^*_t in the simulated distribution, then a solution to (11) and (12) will not always exist.

Fourth, the proof outlined above constructs a distribution of unobservables conditional on (x_s, x_t). The same constructed distribution could be used to exhibit the equivalence of Θ_0 and Θ_S for other "designs," i.e. covariate distributions. For example, given observed combinations of (x_s, x_t) in a particular data set, let $F^e_{y_s,y_t,x_s,x_t}$ denote the corresponding distribution of choices and covariates for this empirical distribution of covariates. Then, $F^e_{y_s,y_t,x_s,x_t}$ can be used to define an identified set Θ^e_0 and a sharp identified set Θ^e_S. By changing only the marginal distribution of covariates and maintaining the same conditional distribution of unobservables to generate $F^e_{y_s,y_t,x_s,x_t}$, Assumption 1 is still satisfied. The conclusion of Theorem 2 follows. Practically, this means that to consider the identifying "power" of alternative covariate designs, the conditional moment inequalities defining Θ_0 still suffice to provide sharp conclusions.

4 Point Identification

In section 3, we showed that the proposed conditional moment inequalities produce a sharp identified set. In this section, we show that under two additional assumptions, a strict inequality version of these conditional moment inequalities yields point identification. In fact, we find the unexpected result that a proper subset of these conditional moment inequalites, based on choice sets of a fixed size, are sufficient to point identify θ_0.

The first additional assumption is the multinomial generalization of Manski (1987)'s additional stochastic assumption for binary choice. This assumption will be used to refine the conditional probability inequalities in Proposition 1.

Assumption 2 *Given the conditioning set (x_s, x_t, λ), the support of the conditional distribution of ε_t is $\mathbb{R}^{\mathscr{D}}$.*

Together with Assumption 1, Assumption 2 ensures that the conditional distributions of the $\mathscr{D}$-dimensional random vectors ε_t and ε_s have supports $\mathbb{R}^{\mathscr{D}}$.

The set inclusion (9) in the proof of Proposition 1 yields the probability inequality in the conclusion of that proposition. Under Assumption 2, when this set inclusion is shown to be a strict inclusion, the probability inequality in the conclusion will also be a strict inequality. A condition leading to strict set inclusion is straightforward to state: If, for some set of choices D and some $d \in D$, $\Delta x_d' \theta > \max_{c \in D^c} \Delta x_c' \theta$, then $\mathcal{E}_{d,D^c,t}$ is a proper subset of $\mathcal{E}_{d,D^c,s}$. This condition leads to the following definition,

$$\widetilde{\mathbb{D}}(x_s, x_t, \theta) = \left\{ D \subset \{0, 1, \ldots, \mathscr{D}\} \,\middle|\, D, D^c \neq \varnothing,\ \min_{d \in D} \Delta x_d' \theta > \max_{c \in D^c} \Delta x_c' \theta \right\}.$$

Now we can state the refinement of Proposition 1 that results from the additional assumption in terms of conditional moment inequalities.

Corollary 3 *Suppose Assumptions 1 and 2 hold.*

(a) Then, for all $D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$,

$$E[m_D(y_s, y_t) \,|\, x_s, x_t] \ > \ 0.$$

(b) If, for some choice set $D \subset \{0, 1, \ldots, \mathscr{D}\}$ with $D, D^c \neq \varnothing$,

$$E[m_D(y_s, y_t) \,|\, x_s, x_t] \ > \ 0,$$

then $D^c \notin \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$.

$\square$

This result extends Proposition 1 by using Assumption 2 to develop conditions for a strict choice probability inequality. This probability inequality is written as a conditional moment inequality in the result above. The conditional moment inequality finding is stated in terms of observables only by integrating out the fixed effects distribution conditional on the covariates, and these conditional moment inequalities based on $\widetilde{\mathbb{D}}(x_s, x_t, \theta)$ generate the following identified set:

$$\widetilde{\Theta}_0 = \{\theta \in \Theta : E[m_D(y_s, y_t) \,|\, x_s, x_t] \; > \; 0 \;\; \forall D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta) \; a.s. \; (x_s, x_t)\}.$$

An additional assumption on covariate variation will lead to the conclusion that $\widetilde{\Theta}_0$ contains only the element θ_0 (up to scale).

The argument behind Corollary 3 is straightforward. For the sets D in the collection $\widetilde{\mathbb{D}}(x_s, x_t, \theta)$ and any $d \in D$, $\mathscr{E}_{d,D^c,t}$ is a proper subset of $\mathscr{E}_{d,D^c,s}$. It follows that the set inclusion in (9) is strict, and so by Assumption 2 the inequality in (8) is strict. Part (a) of this corollary follows by taking the expectation with respect to the distribution of λ conditional on x_s, x_t and restating in terms of conditional moments. The strict inequality in part (a) contrasts with the weak inequality in the conclusion of Proposition 1 and leads to the direct implication of the contrapositive in part (b). The contrapositive result is key to our point identification argument.

It is worthwhile comparing the general multinomial framework above to the special case of binary choice ($\mathscr{D} + 1 = 2$). In the binary choice case, Assumptions 1 and 2 are exactly the stochastic assumptions of Manski (1987). There are three possible orderings for the covariate index difference:

- $\Delta x_1' \theta_0 \; > \; 0 \quad \Longrightarrow \quad \widetilde{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{1\}\}$;
- $\Delta x_1' \theta_0 \; < \; 0 \quad \Longrightarrow \quad \widetilde{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{0\}\}$; and
- $\Delta x_1' \theta_0 \; = \; 0 \quad \Longrightarrow \quad \widetilde{\mathbb{D}}(x_s, x_t, \theta_0) = \varnothing$.

(For reference, $\overline{\mathbb{D}}(x_s, x_t, \theta_0)$ would be the same as $\widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$ in the first two cases, but in the third case $\overline{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{0\}, \{1\}\}$.) By Corollary 3(a),

$$\begin{aligned} \Delta x_1' \theta_0 \; > \; 0 \; &\Longrightarrow \; \widetilde{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{1\}\} \\ &\Longrightarrow \; \Pr(y_s = 1 \,|\, x_s, x_t, \lambda) \; > \; \Pr(y_t = 1 \,|\, x_s, x_t, \lambda) \\ &\Longrightarrow \; E[m_{\{1\}}(y_s, y_t) \,|\, x_s, x_t] \; > \; 0. \end{aligned}$$

Note however that Manski (1987) contains the stronger 'if and only if' conclusion that

$$\Delta x_1' \theta_0 \;>\; 0 \quad \Longleftrightarrow \quad E[m_{\{1\}}(y_s, y_t) \,|\, x_s, x_t] \;>\; 0. \tag{13}$$

This 'if and only if' finding allows Manski (1987) to show point identification by proving sufficient through variation in the covariates. That is, if for some θ, $\Delta x_1' \theta_0 \;>\; 0 \;>\; \Delta x_1' \theta$, or $\Delta x_1' \theta \;>\; 0 \;>\; \Delta x_1' \theta_0$ with positive probability, then $\theta \notin \widetilde{\Theta}_0$. This condition can be used to show that $\theta \notin \widetilde{\Theta}_0$ for all $\theta \neq \theta_0$ (aside from rescalings of θ_0).

Note, however, that an 'if and only if' conclusion, as in (13), does *not* generally hold when $\mathscr{D} + 1 > 2$. To be explicit, by Corollary 3(a),

- $D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0) \implies E[m_D(y_s, y_t) \,|\, x_s, x_t] \;>\; 0,$ but
- $E[m_D(y_s, y_t) \,|\, x_s, x_t] \;>\; 0 \;\not\Longrightarrow\; D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0).$

To give an example, take D to be a set containing exactly one choice, $D = \{d\}$. Then, $E[m_D(y_s, y_t) \,|\, x_s, x_t] \;>\; 0$ means $\Pr(y_s = d | x_s, x_t) > \Pr(y_t = d | x_s, x_t)$. In the binary choice model, Manski (1987)'s Lemma 1 shows that $\{d\} \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$, i.e. $(-1)^{d+1} \Delta x_1' \theta_0 \;>\; 0$. For multinomial choice with greater than two choices, we can*not* conclude that $\{d\} \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$. In other words, we cannot conclude that $d = d^*$, the covariate index difference maximizing choice. Instead, Corollary 3(b) shows that $\{d\}^c \notin \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$. Summarizing, $\Pr(y_s = d | x_s, x_t) > \Pr(y_t = d | x_s, x_t)$ does *not* imply that d is the covariate index difference maximizing choice, but it does rule out that d is the covariate index difference *minimizing* choice. This distinction between the conditional moment inequality implications for the general multinomial choice case and the binary choice case necessitates a new argument for point identification in the multinomial case.

Corollary 3 is used to prove point identification by providing a condition on the covariates that is sufficient to rule out parameter values θ from being an element of the identified set $\widetilde{\Theta}_0$. In particular, if for some θ and some choice set D, $D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$ while $D^c \in \widetilde{\mathbb{D}}(x_s, x_t, \theta)$ with positive probability, then $\theta \notin \widetilde{\Theta}_0$. It follows that the variation in $\widetilde{\mathbb{D}}(x_s, x_t, \theta)$ over covariates and parameters can be used to assess point identification. Next we introduce the assumption on covariate variation that will be used to establish point identification of θ_0.

Suppose the differenced covariate, Δx_d is a K-dimensional vector, so that Δx is a $K\mathscr{D}$-dimensional vector. A superscript will be used to denote each element of the vector, so $\Delta x_d = (\Delta x_d^{[1]}, \ldots, \Delta x_d^{[K]})'$.

Assumption 3

(a) *The support of the distribution of Δx is not contained in any collection of $\mathscr{D}(\mathscr{D}+1)/2$ affine hyperplanes of $\mathbb{R}^{K\mathscr{D}}$.*[6]

(b) *For each choice $d = 1, \ldots, \mathscr{D}$, there is at least one covariate dimension $k_d \in \{1, \ldots, K\}$ such that $\theta_{0,k_d} \neq 0$. Let $\Delta\breve{x} = (\Delta x_1^{[k_1]}, \Delta x_2^{[k_2]}, \ldots, \Delta x_{\mathscr{D}}^{[k_{\mathscr{D}}]})'$. The remaining elements of Δx are collected into a $(K-1)\mathscr{D} \times 1$ vector denoted $\Delta\tilde{x}$. Then, for almost all $\Delta\tilde{x}$, the conditional density of $\Delta\breve{x}$ given $\Delta\tilde{x}$ is everywhere positive on $\mathbb{R}^{\mathscr{D}}$.*

For a given choice set D, consider the region of the Δx-space such that $D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$. Part (a) of Assumption 3 ensures that the variation in Δx lies beyond the boundary of such regions, so that the collection $\widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$ is non-empty with positive probability. Part (b) of the assumption asserts that there is sufficient variation in the covariates with non-zero coefficients, so that the joint density of the true covariate index differences over choices is everywhere positive. For the binary choice case, Assumption 3 is exactly the assumption on covariates in Manski (1987), and Assumption 3 provides a natural generalization to the multinomial choice model.

This assumption is substantive and rules out certain cases. Identification works by ordering covariate indices, $\Delta x_d'\theta$, across choices. Assumption 3(a) ensures that there is sufficient variation across choices for all θ and across time by focusing on Δx_d. Hence, constant terms in the random utility are ruled out, as would be expected given the choice-specific fixed effect specification. As is typical in fixed effects panel models, any covariates that do not vary across time could be effectively subsumed into the fixed effect and their coefficient would not be identified. Hence characteristics of the individual i or choice d that are included as covariates would need to vary across time t to satisfy Assumption 3(a). Assumption 3 also requires across variation across choices in all covariates. That is, characteristics of the individual i that do not vary across choices would be ruled out as a covariate for all choices. Of course, when individual i characteristics are interacted with choice d characteristics (and this interaction varies across time), then the resulting covariate could reasonably satisfy Assumption 3(a). Assumption 3(b) assures unbounded variation in the covariate index differences for all choices. This requirement is expected given that in the binary choice model, Chamberlain (2010) shows that some unboundedness is necessary for binary choice point identification with a general disturbance distribution. Chamberlain's result can be straightforwardly extended to the multinomial choice setting to show that unboundedness is similarly necessary for panel multinomial choice point identification. Assumption 3(b) would fail if each of the covariates for a given choice are bounded or discrete. This assumption, in fact, requires an even stronger form of unbounded variation in the covariates. To satisfy

[6] Affine hyperplanes are hyperplanes that are not required to pass through the origin.

Assumption 3(b), the covariates must have unbounded variation on $\mathbb{R}^{\mathscr{D}}$, and this variation must be achieved through the joint distribution of one covariate for each choice $d = 1, \ldots, \mathscr{D}$. While the restrictions imposed by Assumption 3 are stringent, we have already established sharp identification when Assumption 3 is completely relaxed, placing no restrictions on the covariates.

Now we consider identification of θ_0 under Assumptions 1, 2, and 3. From the multinomial choice setup described in section 2.1, re-scaling random utility does not change the observable variable distributions. In particular, for any $\kappa > 0$, $y_t = y(x_t, \lambda, \varepsilon_t, \theta_0) = y(x_t, \kappa\lambda, \kappa\varepsilon_t, \kappa\theta_0)$, so that identification of θ_0 will only be possible up to scale.

The next result will show that θ_0 is point identified up to scale under Assumptions 1, 2, and 3. That is, $\widetilde{\Theta}_0$ consists only of the single element θ_0 (and positive re-scalings of θ_0). In fact, we obtain a considerably stronger result that point identification can be achieved from a much smaller set of conditional moment inequalities than are used to form $\widetilde{\Theta}_0$. If we fix a value of δ, we can focus on the conditional moment inequalities corresponding to partitions of the choice set into sets of size δ and $\mathscr{D}+1-\delta$, and these inequalities alone will be sufficient to identify θ_0 (up to scale). This conclusion holds for any $\delta \in \{1, 2, \ldots, \mathscr{D}\}$.

For a set of choices D, let $|D|$ denote the size of the set, i.e. the number of choices in D. Then, define

$$\widetilde{\Theta}_{\delta,0} = \{\theta \in \Theta : E[m_D(y_s, y_t) \,|\, x_s, x_t] \,>\, 0 \;\; \forall D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta),\; |D| \in \{\delta, \mathscr{D}{+}1{-}\delta\} \;\; a.s.\, (x_s, x_t)\}.$$

The set $\widetilde{\Theta}_{\delta,0}$ differs from $\widetilde{\Theta}_0$ by using only the moment inequalities generated by the complementary choice set partitions of size, δ and $\mathscr{D}{+}1{-}\delta$. For example, taking $\delta = 1$, $\widetilde{\Theta}_{1,0}$ is based on partitions of the choices into the largest index difference choice (d^*) and its complement along with the smallest covariate index difference choice and its complement. Note that by definition, $\widetilde{\Theta}_0 = \cap_{\delta \in \{1,\ldots,\mathscr{D}\}} \widetilde{\Theta}_{\delta,0}$. Also, by Corollary 3, $\theta_0 \in \widetilde{\Theta}_{\delta,0}$. The next result shows that $\widetilde{\Theta}_{\delta,0}$ consists only of θ_0 (up to scale).

Theorem 4 *Under Assumptions 1, 2, and 3, the conditional moment inequalities defining $\widetilde{\Theta}_{\delta,0}$ identify θ_0 up to scale.*

□

This theorem shows that fixed size partitions of the choice set are sufficient for point identification under the given assumptions. Interestingly, the strict inequality version of the illustrative conditional moment inequality in section 2.2 based on d^*, the maximizing index function difference choice, along with the complementary conditional moment inequality based on the minimizing index function difference choice suffice for point identification of θ_0.

Remarks:

(i) A direct implication of Theorem 4 is that the full set of conditional moment inequalities defining $\widetilde{\Theta}_0$ also suffice to point identify θ_0. Moreover, since a subset of the conditional moment inequalities defining $\widetilde{\Theta}_0$ are sufficient for identification, Theorem 4 shows that θ_0 is over-identified by the use of all conditional moment inequalities defining $\widetilde{\Theta}_0$. Further, Theorem 4 reveals that several separate and non-overlapping subsets of conditional moment inequalities each suffice to achieve identification. The overidentification of θ_0 could be used to form specification tests, which we leave to future work.

(ii) Notice that $\Theta_{\delta,0}$ contains exactly the same complementary set of moment inequalities as $\Theta_{\mathscr{D}+1-\delta,0}$ and so $\Theta_{\delta,0} = \Theta_{\mathscr{D}+1-\delta,0}$. Additionally, when $\mathscr{D}+1$ is even and $\delta = (\mathscr{D}+1)/2$, then $\delta = \mathscr{D} + 1 - \delta$, and the two sets of complementary moment inequalities that define $\Theta_{\delta,0}$ are themselves identical. In this case, $\Theta_{\delta,0}$ is defined solely through the set of inequalities corresponding to choice sets of size δ. A prime example is (panel) binary choice which is point identified by a single conditional moment inequality.

(iii) The proof of point identification does not rely on the manner in which "ties" among random utility maximizing choices are resolved in the multinomial setup. When two or more choices maximize random utility, the setup of section 2.1 assigns the outcome to the largest choice value, but other rules, including randomization, would not affect point identification.

(iv) While the size of the partitioned choice sets is fixed, the choices falling into each partition will vary with the covariates. The above result uses that covariate variation in establishing identification. Also, note that partitioning the choice set (with three or more choices) into two sets does not simplify the multinomial choice model to an equivalent binary choice model with random utility as in (1).

Fix a value $\delta \in \{1, 2, \ldots, \mathscr{D}\}$. Take $\theta \neq a\theta_0$ for any $a \in \mathbb{R}_+$. The proof of Theorem 4 establishes identification by showing that $\theta \notin \widetilde{\Theta}_{\delta,0}$. By Corollary 3, it suffices to prove that for some x, there is a set of choices D such that D contains the choices corresponding to the *largest* δ covariate index differences under θ_0, but D contains the choices corresponding to the *smallest* δ covariate index differences under θ. That is, we can find x such that for some

$|D| = \delta$, $D \in \widetilde{\mathbb{D}}(x_s, x_t, \theta_0)$ and $D^c \in \widetilde{\mathbb{D}}(x_s, x_t, \theta)$. For such a value of x, the corresponding conditional moment inequality evaluated under θ is necessarily violated:

$$\begin{aligned}
& E[m_{D^c}(y_s, y_t) \,|\, x_s, x_t] && (= E\left[\mathbf{1}\{y_s \in D^c\} - \mathbf{1}\{y_t \in D^c\} \,|\, x_s, x_t\right]) \\
& \quad = -E[m_D(y_s, y_t) \,|\, x_s, x_t] && (= -E\left[\mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\} \,|\, x_s, x_t\right]) \\
& \quad < 0 &&
\end{aligned}$$

When such x values exist with positive probability, then $\theta \notin \widetilde{\Theta}_{\delta,0}$.[7]

The conclusion of Theorem 4 is that the conditions on covariates used to achieve point identification in the binary choice case (Manski 1987) can be extended to the multinomial choice case to achieve point identification, and that point identification only relies on a subset of the conditional moment inequalities available.

5 Maximum Score

The sharpness result in section 3 can, of course, be specialized to the case of binary choice. In this case, Theorem 2 provides a (to our knowledge) new supplement to the point identification finding in Manski (1987). In particular, even when point identification fails, the weak version of Manski (1987)'s conditional moment inequalities provide sharp bounds on the binary choice random utility covariate coefficient, θ, under Assumption 1.[8]

This finding raises two interesting questions: When point identification fails in the binary choice case is the maximum score criterion still useful for identifying the sharp identified set? And, would a similar extension of the maximum score criterion be useful in the multinomial case with more than two choices? In the binary choice case, we find a positive answer. Maximum score can be used as a criterion for the sharp identified set even when point identification fails. The analysis in the binary choice case also leads to a natural extension of the maximum score criterion to the general multinomial case. But, here, we find a negative result. The extension of the maximum score criterion does not, generally, yield identification for either point or sharp identification cases.

In the general multinomial choice case, estimation and inference for the identified set, Θ_0, can proceed via conditional moment inequality methods recently developed in Andrews and Shi (2013), Armstrong (2015), Armstrong and Chan (2016), Chernozhukov, Lee, and Rosen (2013), Chetverikov (2018), and Lee, Song, and Whang (2013). These methods provide

[7]This is the multinomial argument analogous to Manski (1987)'s 'relative identification'.

[8]Assumption 1(a) is exactly Manski's Assumption 3. Assumption 1(b) is exactly Manski's Assumption 1(a). So, the sharpness result relaxes Manski's Assumptions 1(b) and 2 (which correspond to Assumptions 2 and 3 in the current work).

valid inference without requiring prior knowledge of whether the identified set is a point. In the point identified binary choice case, Manski (1987) proposes an alternative method of estimation, commonly referred to as maximum score estimation. We first note a connection between the maximum score criterion and the conditional moment inequalities defining the identified set.

The conditional moment inequalities are:

$$E[m_D(y_s, y_t) \,|\, x_s, x_t] \;\geq\; 0, \qquad \forall\; D \in \overline{\mathbb{D}}(x_s, x_t, \theta) \tag{14}$$

where in the binary choice case, as noted in section 2.3,

$$\overline{\mathbb{D}}(x_s, x_t, \theta) = \begin{cases} \{\{1\}\} & \text{for } \Delta x_1'\theta > 0 \\ \{\{0\}\} & \text{for } \Delta x_1'\theta < 0 \\ \{\{0\}, \{1\}\} & \text{for } \Delta x_1'\theta = 0. \end{cases} \tag{15}$$

And, the maximum score criterion function is:

$$H(\theta) = E[\mathrm{sgn}(\Delta x_1'\theta)(y_s - y_t)].$$

To connect these expressions, define a function which "aggregates" the conditional moments across the sets $D \in \overline{\mathbb{D}}(x_s, x_t, \theta)$:

$$H(x_s, x_t, \theta) = \sum_{D \in \overline{\mathbb{D}}(x_s, x_t, \theta)} E[m_D(y_s, y_t) \,|\, x_s, x_t]. \tag{16}$$

Then,

$$H(\theta) = E[H(x_s, x_t, \theta)]. \tag{17}$$

That is, the maximum score criterion is an aggregation of the *un*conditional version of the moments from the conditional moment inequalities.

Clearly, $H(\theta_0) \geq 0$ and $\Theta_0 \subset \{\theta \in \Theta \,|\, H(\theta) \geq 0\}$, but in general this set inclusion would be strict and $\{\theta \in \Theta \,|\, H(\theta) \geq 0\}$ would not be sharp. Instead of checking non-negativity of this criterion, maximum score seeks to maximize it. When Assumptions 1, 2, and 3 hold, point identification is achieved, and, according to Manski (1987) in the binary choice case, the maximum score criterion $H(\theta)$ is uniquely maximized at θ_0.

The following proposition shows that, in the binary choice case, the maximum score criterion is useful even when point identification is not achieved. In particular, under Assumption 1 alone, Θ_0 could be either a set or a point, and the maximum score criterion $H(\theta)$ exactly identifies this set (or point) Θ_0.

Proposition 5 *Suppose $\mathscr{D}+1=2$ (binary choice) and Assumption 1 holds. Then,*

$$\Theta_0 \;=\; \arg\max_{\theta\in\Theta} H(\theta) \;=\; \{\theta\in\Theta : H(\theta)=H(\theta_0)\}.$$

$\square$

Under the conditions of Theorem 2, Θ_0 is itself sharp, so the maximum score criterion will identify the sharp bounds on the parameter.

As noted in the discussion following Corollary 3, the conditional moment inequality implications are distinctly different for binary choice and multinomial choice with at least three choices. This distinction is also apparent from when we seek to extend the binary choice maximum score notion to the general multinomial choice case. The criterion function $H(\theta)$ given in (16) and (17) is still well-defined the general multinomial choice case, where the set $\overline{\mathbb{D}}(x_s, x_t, \theta)$ is defined as in (7) rather than as given in the binary choice case in (15). However, the identification results based on maximizing $H(\theta)$ do not extend to the multinomial case with three or more choices. In particular, we can show by counterexample (see Appendix) that under Assumptions 1, 2, and 3, though θ_0 is point identified by Theorem 4, θ_0 may not be a maximizer of $H(\theta)$. Further, under only Assumption 1, where Θ_0 is the sharp identified set, again the criterion $H(\theta)$ may not be maximized at θ_0. So, an extension of Proposition 5 to the general multinomial choice model is not generally possible.

This section shows that the maximum score criterion provides a robust identification tool for binary choice. For the extension to the multinomial choice case given here, the criterion cannot be relied upon for identification. This extension does fit with the original motivation provided in binary choice ("score" maximization) and is a simple aggregation of conditional moments. However, other aggregations could potentially be used to provide alternative maximum score criterion definitions in the multinomial case. Note that as long as these alternative criterion definitions place positive weights on the unconditional moments m_D for all $D \in \overline{\mathbb{D}}(x_s, x_t, \theta)$ and all (x_s, x_t), then the conclusions of this section will still hold.

6 Conclusion

We have provided a new approach to identification for multinomial choice models. Our focus has been on the multinomial choice model which allows for choice-specific fixed effects with a group (or panel) structure and a nonparametric distribution of disturbances only restricted to satisfy a stationarity assumption. We show that this specification generates conditional moment inequalities which can be used for identification of the parameter vector and avoids the incidental parameter problem using only two time periods. These conditional moment inequalities point identify the coefficient parameter under certain conditions on the covariates, and provide sharp bounds when the covariate conditions are relaxed. Here we point to some avenues for extensions and future research.

One conclusion from our identification findings is that robust methods of estimation and inference using our conditional moment inequalities have the virtue of yielding sharp bounds whether or not point identification holds. However, if point identification is known to hold, then one could pursue specialized estimation and inference methods for this case. From section 5, the maximum score criterion will not generally work for this purposes, but a criterion that penalizes violations of the conditional moment inequalities is possible. Aradillas-López and Rosen (2016) and Kahn, Ouyang, and Tamer (2019) have proposed estimation methods for point identified conditional moment inequality models, but these papers focus on cases where the contact set has positive measure and parametric rate consistency is possible. The multinomial choice model will generally only have a measure zero contact set, so these methods would need to be extended to be applied in the current setting.

Often the focus of empirical studies is not on θ_0 per se but rather on different functionals that could depend on θ_0 (e.g. Tebaldi, Torgovitsky, and Yang 2018). In the discrete choice panel data setting, Chernozhukov, Fernández-Val, Hahn, and Newey (2013) suggest particular functionals of interest such as the conditional quantile or average structural effects. In such cases, our conditional moment inequalities provide a new source of identifying information. Without restricting the disturbance distribution across choices, our conditional moment inequalities are relatively easy to compute and provide sharp (and sometimes point) identifying information on θ_0. This additional "within" information can be used to improve upon the estimation of the various effects by narrowing the range of parameter values to be considered together with the possible disturbance distributions (including fixed effects) that are consistent with the "between" variation specified in the Chernozhukov, Fernández-Val, Hahn, and Newey (2013) paper.

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.

Ahn, H., H. Ichimura, J. Powell, and P. Ruud (2017): "Simple Estimators for Invertible Index Models," *Journal of Business Economics and Statistics*, 36(1), 1–10.

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

Aradillas-López, A., and A. Rosen (2016): "Inference in Ordered Response Games with Complete Information," CeMMAP Working Paper.

Armstrong, T. (2015): "Asymptotically Exact Inference in Conditional Moment Inequality Models," *Journal of Econometrics*, 186(1), 51–65.

Armstrong, T., and H. P. Chan (2016): "Multiscale Adaptive Inference on Conditional Moment Inequalities," *Journal of Econometrics*, 194(1), 24–43.

Berry, S., J. Levinsohn, and A. Pakes (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.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(21), 159–168.

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. (2018): "Adaptive Test of Conditional Moment Inequalities," *Econometric Theory*, 34(1), 186–227.

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

Gao, W., and M. Li (2018): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," Yale Working Paper.

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

Ho, K., and A. Pakes (2014): "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.

Kahn, S., F. Ouyang, and E. Tamer (2019): "Inference on Semiparametric Multinomial Response Models," Harvard University Working Paper.

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

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing Functional Inequalities," *Journal of Econometrics*, 172(1), 14–32.

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.

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.

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

Pakes, A., and J. Porter (2014): "Moment Inequalities for Multinomial Choice with Fixed Effects," University of Wisconsin Working Paper.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015): "Moment Inequalities and Their Application," *Econometrica*, 80(4), 315–334.

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

Shi, X., M. Shum, and W. Song (2018): "Estimating Semiparametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86(2), 737–761.

Tebaldi, P., A. Torgovitsky, and H. Yang (2018): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," University of Chicago Working Paper.

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

7 Appendix

Proof of Theorem 2: Take $F_{y_s,y_t,x_s,x_t} \in \mathscr{F}_{ob}$. It is straightforward to show that $\Theta_S \subset \Theta_0$. So we will focus on showing $\Theta_0 \subset \Theta_S$. Take $\theta \in \Theta_0$. For all x_s, x_t in the support of the joint covariate space, we will exhibit a conditional distribution $(\varepsilon_s^*, \varepsilon_t^*)|x_s, x_t$ satisfying Assumption 1(b) with $\lambda^* = 0$ and $F_{y_s^*, y_t^* \,|\, x_s, x_t} = F_{y_s, y_t \,|\, x_s, x_t}$ where $y_j^* = y(x_j, \lambda^* = 0, \varepsilon_j^*, \theta)$ for $j = s, t$.

Suppose we order the covariate indices for the parameter θ and there is a strict ordering: $\Delta x'_{(\mathscr{D})}\theta > \Delta x'_{(\mathscr{D}-1)}\theta > \cdots > \Delta x'_{(0)}\theta$ (using order statistic subscript notation). Since $\theta \in \Theta_0$, the conditional moment inequalities imply

$$\Pr(y_s \in \{(\mathscr{D}), \ldots, (d)\} \,|\, x_s, x_t) \geq \Pr(y_t \in \{(\mathscr{D}), \ldots, (d)\} \,|\, x_s, x_t) \quad \forall d = 1, 2, \ldots, \mathscr{D}.$$

Let $p_{d,d'} = \Pr(y_s = d, y_t = d'|x_s, x_t)$ and $p^*_{d,d'} = \Pr(y_s^* = d, y_t^* = d'|x_s, x_t)$. We need to find $(\varepsilon_s^*, \varepsilon_t^*)|x_s, x_t$ such that $p^*_{d,d'} = p_{d,d'}$ $\forall d, d'$ and $\varepsilon_s^*|x_s, x_t \;\sim\; \varepsilon_t^*|x_s, x_t$.

Define $R_{d;s} = \{\varepsilon : y(x_s, \lambda^* = 0, \varepsilon_s^*, \theta) = d\}$. The set inclusion obtained in the proof of Proposition 1 shows that

$$R_{(\mathscr{D});t} \cup \cdots \cup R_{(d);t} \;\subset\; R_{(\mathscr{D});s} \cup \cdots \cup R_{(d);s}\ , \quad \forall d \in \{1, \ldots, \mathscr{D}\}. \tag{18}$$

Since the sets $R_{(d);s}$ form a partition for $d = 0, \ldots, \mathscr{D}$, the set inclusion (18) implies that

$$R_{(d);s} \,\cap\, R_{(d');t} = \emptyset \quad \text{for } d' > d. \tag{19}$$

Let $R_{d,d'} = R_{d;s} \cap R_{d';t}$, which is a set in the ε_s^*-space (or the ε_t^*-space). Cartesian products of these sets will form sets in the $(\varepsilon_s^*, \varepsilon_t^*)$-space: let $R_{d,d' \times d'',d'''} = R_{d,d'} \times R_{d'',d'''} = \{(\varepsilon_s^*, \varepsilon_t^*) : \varepsilon_s^* \in R_{d,d'}, \varepsilon_t^* \in R_{d'',d'''}\}$. Finally, let $q^*_{d,d' \times d'',d'''} = \Pr((\varepsilon_s^*, \varepsilon_t^*) \in R_{d,d' \times d'',d'''}|x_s, x_t)$. These probabilities form the basic building blocks for our constructed $(\varepsilon_s^*, \varepsilon_t^*)|x_s, x_t$ distribution, as $R_{d,d' \times d'',d'''}$ partitions the $(\varepsilon_s^*, \varepsilon_t^*)$-space. Given (19),

$$p^*_{(d),(d')} \;=\; \sum_{\underset{\sim}{d}=0}^{\mathscr{D}} \sum_{\tilde{d}=0}^{\mathscr{D}} q^*_{(d),(\underset{\sim}{d}) \times (\tilde{d}),(d')} \;=\; \sum_{\underset{\sim}{d}=0}^{d} \sum_{\tilde{d}=d'}^{\mathscr{D}} q^*_{(d),(\underset{\sim}{d}) \times (\tilde{d}),(d')}.$$

To get the constructed distribution to match the observed distribution, we will need to show that there exists $q^*_{d,d' \times d'',d'''}$ satisfying

$$p_{(d),(d')} \;=\; \sum_{\underset{\sim}{d}=0}^{d} \sum_{\tilde{d}=d'}^{\mathscr{D}} q^*_{(d),(\underset{\sim}{d}) \times (\tilde{d}),(d')}, \tag{20}$$

as well as ensuring that Assumption 1(b) holds for the contructed distribution. For each $R_{d,d'} \neq \emptyset$, choose a point $r_{d,d'} \in R_{d,d'}$. Define $(\varepsilon_s^*, \varepsilon_t^*)|x_s, x_t$ to be the discrete distribution on the support points $(r_{d,d'}, r_{d'',d'''})$, $\Pr((\varepsilon_s^*, \varepsilon_t^*) = (r_{d,d'}, r_{d'',d'''})|x_s, x_t) = q^*_{d,d'\times d'',d'''}$. So the marginal is

$$\Pr(\varepsilon_s^* = r_{(d),(d')}|x_s, x_t) \;=\; \sum_{\underset{\sim}{d}=0}^{\mathscr{D}} \sum_{\tilde{d}=\underset{\sim}{d}}^{\mathscr{D}} q^*_{(d),(d')\times(\tilde{d}),(\underset{\sim}{d})}.$$

The marginal for $\varepsilon_t^*|x_s, x_t$ is similar. To ensure that Assumption 1(b) is satisfied we will need for the marginals to match, for $d \geq d'$,

$$0 \;=\; \sum_{\underset{\sim}{d} \leq \tilde{d}} \left(q^*_{(\tilde{d}),(\underset{\sim}{d})\times(d),(d')} - q^*_{(d),(d')\times(\tilde{d}),(\underset{\sim}{d})} \right). \tag{21}$$

In addition to equations (20) and (21), the nonegativity inequalities $q^*_{d,d'\times d'',d'''} \geq 0$ must hold. Let p denote the vector of joint probabilities, $p = (p_{(\mathscr{D}),(\mathscr{D})}, p_{(\mathscr{D}),(\mathscr{D}-1)}, \ldots)'$. Let q^* be the vector of probabilities $q^*_{(d),(d')\times(d''),(d''')}$ (where $d \geq d'$ and $d'' \geq d'''$), $q^* = (q^*_{(\mathscr{D}),(\mathscr{D})\times(\mathscr{D}),(\mathscr{D})}, \ldots)'$. And let Q_s be the matrix with elements in $\{0, 1\}$ such that equation (20) can be restated as $p = Q_s q^*$, and let Q_p be the matrix with elements in $\{-1, 0, 1\}$ such that equation (21) can be restated as $0 = Q_p q^*$.

Our goal then can be summarized as showing that $\exists q^* \geq 0$ such that: (A) $p = Q_s q^*$; and (B) $0 = Q_p q^*$. Let z be a $(\mathscr{D}+1)^2$-dimensional vector conformable with p, $z = (z_{(\mathscr{D}),(\mathscr{D})}, z_{(\mathscr{D}),(\mathscr{D}-1)}, \ldots)'$. Let w be a $(\mathscr{D}+1)(\mathscr{D}+2)/2$-dimensional vector, $w = (\ldots, w_{(d),(d')}, \ldots)'$. Farkas' Lemma states that if

$$\begin{pmatrix} z \\ w \end{pmatrix}' \begin{pmatrix} Q_s \\ Q_p \end{pmatrix} \geq 0 \;\text{ implies }\; \begin{pmatrix} z \\ w \end{pmatrix}' \begin{pmatrix} p \\ 0 \end{pmatrix} = z'p \geq 0,$$

then $\exists q^* \geq 0$ satisfying (A) and (B) above.

Each element $q^*_{(d),(d')\times(d''),(d''')}$ of q^* appears in exactly one equation from constraints (A) and either zero or two (with positive and negative signs) from constraints (B). In particular, elements of the form $q^*_{(d),(d)\times(d),(d)}$ (and $q^*_{(d),(d')\times(d),(d')}$ with $d > d'$) appear in (A) but not (B). Hence, $z_{(d),(d)} \geq 0$ (and $z_{(d),(d')} \geq 0$ for $d > d'$). Also, the conditional moment inequalities yield that for $d = 1, \ldots, \mathscr{D}$, $\Pr(y_s \in \{(\mathscr{D}), \ldots, (d)\}|x_s, x_t) \geq \Pr(y_t \in \{(\mathscr{D}), \ldots, (d)\}|x_s, x_t)$ which implies

$$\sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\tilde{d}),(\underset{\sim}{d})} \;\geq\; \sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\underset{\sim}{d}),(\tilde{d})} \quad \text{for } d = 1, \ldots, \mathscr{D} \tag{22}$$

Now, for constants a_d,

$$\begin{aligned}
z'p &= \sum_{d=0}^{\mathscr{D}} z_{(d),(d)} p_{(d),(d)} \ + \sum_{d>d'} \left(z_{(d),(d')} p_{(d),(d')} + z_{(d'),(d)} p_{(d'),(d)} \right) \\
&\geq \sum_{d>d'} \left(z_{(d),(d')} p_{(d),(d')} + z_{(d'),(d)} p_{(d'),(d)} \right) \\
&= \sum_{d=1}^{\mathscr{D}} a_d \underbrace{\left[\sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\tilde{d}),(\underset{\sim}{d})} - \sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\underset{\sim}{d}),(\tilde{d})} \right]}_{\geq 0 \text{ by } (22)} \\
&\quad + \sum_{s=1}^{\mathscr{D}} \sum_{d=0}^{\mathscr{D}-s} \left\{ \left(z_{(d+s),(d)} - \left[\sum_{d'=d+1}^{d+s} a_{d'} \right] \right) p_{(d+s),(d)} + \left(z_{(d),(d+s)} - \left[\sum_{d'=d+1}^{d+s} a_{d'} \right] \right) p_{(d),(d+s)} \right\}
\end{aligned}$$

So, given $\begin{pmatrix} z \\ w \end{pmatrix}' \begin{pmatrix} Q_s \\ Q_p \end{pmatrix} \geq 0$, we have $z'p \geq 0$ if $\exists\, a_{\mathscr{D}}, \ldots, a_1 \geq 0$ such that $-z_{(d),(d+s)} \leq \sum_{d'=d+1}^{d+s} a_{d'} \leq z_{(d+s),(d)}$ for $s = 1, \ldots, \mathscr{D}$, $d = 0, \ldots, \mathscr{D} - s$.

From examination of the constraints (A) and (B), $\begin{pmatrix} z \\ w \end{pmatrix}' \begin{pmatrix} Q_s \\ Q_p \end{pmatrix} \geq 0$ yields, for $d \neq d'$, $z_{(d),(d')} + w_{(\tilde{d}),(d')} - w_{(d),(\underset{\sim}{d})} \geq 0$ with $\tilde{d} \in \{d', \ldots, \mathscr{D}\}, \underset{\sim}{d} \in \{0, \ldots, d\}$. For $d > d'$, let $\bar{a}_{d,d'+1} = \max_{\tilde{d} \geq d', \underset{\sim}{d} \leq d} \left\{ w_{(d),(\underset{\sim}{d})} - w_{(\tilde{d}),(d')} \right\}$ (and we will shorten $\bar{a}_{d,d}$ to $\bar{a}_d$). For $d+1 < d'$, let $\underset{\sim}{a}_{d',d+1} = \min_{\tilde{d} \geq d', \underset{\sim}{d} \leq d} \left\{ w_{(\tilde{d}),(d')} - w_{(d),(\underset{\sim}{d})} \right\}$. For $d + 1 = d'$, let $\underset{\sim}{a}_{d'} = \max \left\{ 0, \min_{\tilde{d} \geq d', \underset{\sim}{d} \leq d'-1} \left\{ w_{(\tilde{d}),(d')} - w_{(d'-1),(\underset{\sim}{d})} \right\} \right\}$. Then, for $d > d'$, $z_{(d),(d')} \geq \bar{a}_{d,d'+1}$. And, for $d < d'$, $-z_{(d),(d')} \leq \underset{\sim}{a}_{d',d+1}$, where we have imposed (22) through the definition of $\underset{\sim}{a}_{d'}$. So, to show $\exists\, a_{\mathscr{D}}, \ldots, a_1 \geq 0$ such that $-z_{(d),(d+s)} \leq \sum_{d'=d+1}^{d+s} a_{d'} \leq z_{(d+s),(d)}$ for $s = 1, \ldots, \mathscr{D}$, $d = 0, \ldots, \mathscr{D} - s$, it will suffice to show $\exists\, a_{\mathscr{D}}, \ldots, a_1$ such that $\underset{\sim}{a}_d \leq a_d \leq \bar{a}_d$ for $d = 1, \ldots, \mathscr{D}$, and $\underset{\sim}{a}_{d+s,d+1} \leq \sum_{d'=d+1}^{d+s} a_{d'} \leq \bar{a}_{d+s,d+1}$ for $s = 1, \ldots, \mathscr{D}$, $d = 0, \ldots, \mathscr{D} - s$. We can show this by proving that certain linear combinations of the lower bounds are smaller than certain linear combinations of the upper bounds. Let $b_{\cdot}$ and $c_{\cdot}$ denote the coefficients in the linear combinations. In particular, it is sufficient to show that

$$\sum_{d=1}^{\mathscr{D}} b_d \underset{\sim}{a}_d \ + \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} b_{d',d} \underset{\sim}{a}_{d',d} \leq \sum_{d=1}^{\mathscr{D}} c_d \bar{a}_d \ + \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} c_{d',d} \bar{a}_{d',d} \tag{23}$$

where $b_{\cdot} \geq 0$ and $c_{\cdot} \geq 0$ and for $d = 1, \ldots, \mathscr{D}$,

$$b_d + \sum_{1 \leq d' \leq d \leq d'' \leq \mathscr{D}} b_{d'',d'} \ = \ c_d + \sum_{1 \leq d' \leq d \leq d'' \leq \mathscr{D}} c_{d'',d'}. \tag{24}$$

Before showing (23), it will be useful to first note a fact about the coefficients $b.$ and $c.$. Let $g_d = \min\{b_d, c_d\}$ and $g_{d',d} = \min\{b_{d',d}, c_{d',d}\}$ for all d', d. Equation (23) can be re-stated as

$$\sum_{d=1}^{\mathscr{D}}(b_d - g_d)\underline{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}}(b_{d',d} - g_{d',d})\underline{a}_{d',d} + \left[\sum_{d=1}^{\mathscr{D}} g_d\underline{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}} g_{d',d}\underline{a}_{d',d}\right]$$
$$\leq \; \sum_{d=1}^{\mathscr{D}}(c_d - g_d)\bar{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}}(c_{d',d} - g_{d',d})\bar{a}_{d',d} + \left[\sum_{d=1}^{\mathscr{D}} g_d\bar{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}} g_{d',d}\bar{a}_{d',d}\right].$$

Since $\underline{a}_d \leq \bar{a}_d$ and $\underline{a}_{d',d} \leq \bar{a}_{d',d}$

$$\sum_{d=1}^{\mathscr{D}} g_d\underline{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}} g_{d',d}\underline{a}_{d',d} \; \leq \; \sum_{d=1}^{\mathscr{D}} g_d\bar{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}} g_{d',d}\bar{a}_{d',d}.$$

So it will suffice to show

$$\sum_{d=1}^{\mathscr{D}}(b_d - g_d)\underline{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}}(b_{d',d} - g_{d',d})\underline{a}_{d',d} \; \leq \; \sum_{d=1}^{\mathscr{D}}(c_d - g_d)\bar{a}_d \; + \; \sum_{d=1}^{\mathscr{D}-1}\sum_{d'=d+1}^{\mathscr{D}}(c_{d',d} - g_{d',d})\bar{a}_{d',d}.$$

Or, more simply, in proving (23), we may assume either $b_d = 0$ or $c_d = 0$ holds for each d, and similarly $b_{d',d} = 0$ or $c_{d',d} = 0$ for each d', d, which is useful in the cases to be considered.

Take the case where $\underline{a}_d > 0 \; \forall d$, the argument for (23) follows.

$$\begin{aligned}
&\sum_{d=1}^{\mathscr{D}} c_d \bar{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} c_{d',d} \bar{a}_{d',d} \\
&\geq \; \sum_{d=1}^{\mathscr{D}} c_d (w_{d,d} - w_{d-1,d-1}) \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} c_{d',d} (w_{d',d'} - w_{d-1,d-1}) \\
&= \; w_{\mathscr{D},\mathscr{D}} \left[c_{\mathscr{D}} \;+\; \sum_{d=1}^{\mathscr{D}-1} c_{\mathscr{D},d} \right] \;+\; w_{\mathscr{D}-1,\mathscr{D}-1} \left[(c_{\mathscr{D}-1} - c_{\mathscr{D}}) \;+\; \sum_{d'=1}^{\mathscr{D}-2} c_{\mathscr{D}-1,d'} \right] \\
&\quad + \sum_{d=2}^{\mathscr{D}-2} w_{d,d} \left[(c_d - c_{d+1}) \;+\; \sum_{d'=1}^{d-1} c_{d,d'} \;-\; \sum_{d'=d+2}^{\mathscr{D}} c_{d',d+1} \right] \\
&\quad + w_{1,1} \left[(c_1 - c_2) \;-\; \sum_{d'=3}^{\mathscr{D}} c_{d',2} \right] \;+\; w_{0,0} \left[-c_1 \;-\; \sum_{d'=2}^{\mathscr{D}} c_{d',1} \right] \\
&= \; w_{\mathscr{D},\mathscr{D}} \left[b_{\mathscr{D}} \;+\; \sum_{d=1}^{\mathscr{D}-1} b_{\mathscr{D},d} \right] \;+\; w_{\mathscr{D}-1,\mathscr{D}-1} \left[(b_{\mathscr{D}-1} - b_{\mathscr{D}}) \;+\; \sum_{d'=1}^{\mathscr{D}-2} b_{\mathscr{D}-1,d'} \right] \\
&\quad + \sum_{d=2}^{\mathscr{D}-2} w_{d,d} \left[(b_d - b_{d+1}) \;+\; \sum_{d'=1}^{d-1} b_{d,d'} \;-\; \sum_{d'=d+2}^{\mathscr{D}} b_{d',d+1} \right] \\
&\quad + w_{1,1} \left[(b_1 - b_2) \;-\; \sum_{d'=3}^{\mathscr{D}} b_{d',2} \right] \;+\; w_{0,0} \left[-b_1 \;-\; \sum_{d'=2}^{\mathscr{D}} b_{d',1} \right] \\
&= \; \sum_{d=1}^{\mathscr{D}} b_d (w_{d,d} - w_{d-1,d-1}) \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} b_{d',d} (w_{d',d'} - w_{d-1,d-1}) \\
&\geq \; \sum_{d=1}^{\mathscr{D}} c_d \underline{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} c_{d',d} \underline{a}_{d',d}
\end{aligned}$$

where the second equality follows from differences of the equations given in (24).

The other cases to check allow $\underline{a}_d = 0$ for some d and work similarly. We have focused on the case where there is a strict covariate index ordering. When the covariate index ordering is weak (includes some ties), the argument above simplifies. Finally, having verified Farkas' Lemma, we can conclude that a constructed disturbance distribution exists that satisfies Assumption 1 and generates a constructed outcome and covariate distribution that matches the observed distribution. It follows that $\Theta_0 \subset \Theta_S$ and Θ_0 is sharp. □

Proof of Theorem 4: Given θ_0, we will show that θ_0 is identified relative to θ for any θ such that $\theta = 0$ or $\theta_0/\|\theta_0\| \neq \theta/\|\theta\|$. Fix an integer δ with $1 \leq \delta \leq \mathscr{D}$. We will show that partitioning the the choice set into subsets of size δ and $\mathscr{D} + 1 - \delta$ will be sufficient to identify θ_0.

We divide the argument into three cases. We want to identify θ_0 relative to $\theta = -\theta_0$, $\theta = 0$, and $\theta \neq c\theta_0$. The first two cases are straightforward. The key parts of the argument are contained in the last case. It will be useful to define the following sets. Given a pair of choices $d, d' \in \{0, 1, \ldots, \mathscr{D}\}$, let

$$S_{d,d'}(\theta) \quad = \quad \{\, \Delta x \,:\, \Delta x_d'\theta = \Delta x_{d'}'\theta \,\}.$$

Given a partition of choices D and D^c, let

$$A_{D,D^c}(\theta) \;=\; \{\, \Delta x \,:\, \min_{d\in D} \Delta x_d'\theta > \max_{d'\in D^c} \Delta x_{d'}'\theta \,\}$$

Case 1. $\frac{\theta}{\|\theta\|} = -\frac{\theta_0}{\|\theta_0\|}$.

$$A_{D^c,D}(\theta) = \{\Delta x :\; \min_{d'\in D^c} \Delta x_{d'}'\theta \;>\; \max_{d\in D} \Delta x_d'\theta\} = \left\{\Delta x :\; \min_{d'\in D^c} \Delta x_{d'}'\theta\frac{\|\theta_0\|}{\|\theta\|} \;>\; \max_{d\in D} \Delta x_d'\theta\frac{\|\theta_0\|}{\|\theta\|}\right\}$$

$$= \{\Delta x :\; \min_{d'\in D^c} -\Delta x_{d'}'\theta_0 \;>\; \max_{d\in D} -\Delta x_d'\theta_0\} = \{\Delta x :\; \min_{d\in D} \Delta x_d'\theta_0 \;>\; \max_{d'\in D^c} \Delta x_{d'}'\theta_0\} = A_{D,D^c}(\theta_0)$$

Hence, $\sum_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} \Pr(A_{D,D^c}(\theta_0) \cap A_{D^c,D}(\theta)) = \sum_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} \Pr(A_{D,D^c}(\theta_0))$. Suppose $\sum_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} \Pr(A_{D,D^c}(\theta_0)) = 0$. Since $\left(\cup_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} A_{D,D^c}(\theta_0)\right)^c \subseteq \cup_{d\neq d'} S_{d,d'}(\theta_0)$, we must then have $1 = \Pr\left(\cup_{d\neq d'} S_{d,d'}(\theta_0)\right)$. But, $\cup_{d\neq d'} S_{d,d'}(\theta_0)$ is a collection of $\mathscr{D}(\mathscr{D}+1)/2$ hyperplanes contradicting Assumption 3(a), which means that our supposition is wrong and we have $\sum_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} \Pr(A_{D,D^c}(\theta_0) \cap A_{D^c,D}(\theta)) > 0$.

Case 2. $\theta = 0$.

From Case 1, $\sum_{D\subset\{0,1,\ldots,\mathscr{D}\}:|D|=\delta} \Pr(A_{D,D^c}(\theta_0)) > 0$, so $\Pr(A_{D,D^c}(\theta_0)) > 0$ for some D with $|D| = \delta$. By Corollary 3(a), for x_s, x_t such that $\Delta x \in A_{D,D^c}(\theta_0)$, we have $\Pr(y_s \in D \,|\, x_s, x_t) > \Pr(y_t \in D \,|\, x_s, x_t)$. So, $\Pr(\{\Delta x : \Pr(y_s \in D \,|\, x_s, x_t) > \Pr(y_t \in D \,|\, x_s, x_t)\}) > 0$, which implies that θ_0 is identified relative to $\theta = 0$.

Case 3. $\theta \neq c\theta_0$ for any $c \in \mathbb{R}$.

Given Assumption 3(b) and a set $A_{D,D^c}(\theta)$ for any value of θ, it will be useful to consider the projection of the $A_{D,D^c}(\theta)$ onto the $\Delta\breve{x}$ space conditioning on a value of $\Delta\tilde{x}$:

$$\breve{A}_{D,D^c}(\theta|\Delta\tilde{x}) = \{\, \Delta\breve{x} \,:\Delta x = (\Delta\breve{x}, \Delta\tilde{x}) \in A_{D,D^c}(\theta)\} = \{\, \Delta\breve{x} \,:\, \min_{d\in D} \Delta x_d'\theta > \max_{d'\in D^c} \Delta x_{d'}'\theta \,\}.$$

We can similarly project the sets $S_{d,d'}(\theta)$:

$$\breve{S}_{d,d'}(\theta|\Delta\tilde{x}) = \{\, \Delta\breve{x} \,:\Delta x_i = (\Delta\breve{x}, \Delta\tilde{x}) \in S_{d,d'}(\theta)\} = \{\, \Delta\breve{x} \,:\, \Delta x_d'\theta = \Delta x_{d'}'\theta \,\}.$$

Let ∂A denote the boundary of set A, and let $\breve{S}(\theta|\Delta\tilde{x}) = \bigcup_{d\neq d'} \breve{S}_{d,d'}(\theta|\Delta\tilde{x})$. Note that $\bigcup_{D:|D|=\delta} \partial\breve{A}_{D^c,D}(\theta|\Delta\tilde{x}) \subseteq \breve{S}(\theta|\Delta\tilde{x})$ so that $\left(\bigcup_{D:|D|=\delta} \breve{A}_{D^c,D}(\theta|\Delta\tilde{x})\right) \cup \breve{S}(\theta|\Delta\tilde{x}) = \mathbb{R}^{\mathscr{D}}$.

The hyperplanes $\breve{S}_{d,d'}(\theta_0|\Delta\tilde{x})$ have a unique point of intersection, which we will denote by $\breve{c}$. That is, fixing $\Delta\tilde{x}$, $\breve{c}$ is the unique solution to $\Delta x_d'\theta_0 = \Delta x_{d'}'\theta_0$ for all $d \neq d'$:

$$\breve{c} \;=\; \bigcap_{d\neq d'} \breve{S}_{d,d'}(\theta_0|\Delta\tilde{x}) \;=\; \left(-\Delta\tilde{x}_1'\left(\frac{\tilde{\theta}_{0,1}}{\breve{\theta}_{0,1}}\right), \ldots, -\Delta\tilde{x}_{\mathscr{D}}'\left(\frac{\tilde{\theta}_{0,\mathscr{D}}}{\breve{\theta}_{0,\mathscr{D}}}\right)\right).$$

Note that $\breve{c} \in \partial\breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x})$ for all D, and $\breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x})$ is a non-empty open set in $\mathbb{R}^{\mathscr{D}}$. Then, for any open ball $\mathscr{B}$ centered at $\breve{c}$, by Assumption 3(b), $\Pr(\mathscr{B} \cap \breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x}) \,|\, \Delta\tilde{x}) > 0$.

Suppose, for the moment, that $\breve{c} \notin \breve{S}(\theta|\Delta\tilde{x})$. Then, for some D, $\breve{c} \in \breve{A}_{D^c,D}(\theta|\Delta\tilde{x})$. Since, $\breve{A}_{D^c,D}(\theta|\Delta\tilde{x})$ is open, there exists some ball $\mathscr{B}$ centered at $\breve{c}$ such that $\mathscr{B} \subset \breve{A}_{D^c,D}(\theta|\Delta\tilde{x})$. As noted above, $\mathscr{B} \cap \breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x})$ is a non-empty open set, and so

$$\Pr(\breve{A}_{D^c,D}(\theta|\Delta\tilde{x}) \cap \breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x}) \,|\, \Delta\tilde{x}) \geq \Pr(\mathscr{B} \cap \breve{A}_{D,D^c}(\theta_0|\Delta\tilde{x}) \,|\, \Delta\tilde{x}) \;>\; 0.$$

Our conclusion will follow if we can show that $\Pr\left(\breve{c} \notin \breve{S}(\theta|\Delta\tilde{x})\right) > 0$.

Now suppose $\breve{c} \in \breve{S}(\theta|\Delta\tilde{x})$. Then, there is some d, d' such that $\breve{c} \in \breve{S}_{d,d'}(\theta|\Delta\tilde{x})$ and hence

$$\breve{c}_d\breve{\theta}_d + \Delta\tilde{x}_d'\tilde{\theta}_d \;=\; \breve{c}_{d'}\breve{\theta}_{d'} + \Delta\tilde{x}_{d'}'\tilde{\theta}_{d'} \quad\Longleftrightarrow\quad \Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) \;=\; \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right).$$

Suppose $\tilde{\theta}_d = \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)$ to arrive at a contradiction.

a) $\underline{\breve{\theta}_d = 0}$: Then $\tilde{\theta}_d = 0$, which would imply $\theta = 0$. This contradicts our supposition, so we cannot have $\breve{\theta}_d = 0$.

b) $\underline{\breve{\theta}_d \neq 0}$: Then, $\frac{\tilde{\theta}_d}{\breve{\theta}_d} = \frac{\tilde{\theta}_{0,d}}{\breve{\theta}_{0,d}} \implies \theta_0 = \left(\breve{\theta}_{0,d}\,,\, \breve{\theta}_{0,d}\cdot\frac{\tilde{\theta}_{0,d}}{\breve{\theta}_{0,d}}\right) = \left(\breve{\theta}_{0,d}\,,\, \breve{\theta}_{0,d}\cdot\frac{\tilde{\theta}_d}{\breve{\theta}_d}\right) = \frac{\breve{\theta}_{0,d}}{\breve{\theta}_d}\cdot\theta$, but this conclusion contradicts our supposition that $\theta \neq c\theta_0$.

It follows that $\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right) \neq 0$, and similarly that $\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right) \neq 0$. So,

$$\left\{\breve{c} \in \breve{S}(\theta|\Delta\tilde{x})\right\} \subseteq \left\{\Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) = \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right) \text{for some } d \neq d'\right\}$$

Suppose $\Pr\left(\breve{c}(\Delta\tilde{x}) \in \breve{S}(\theta|\Delta\tilde{x})\right) = 1$ to arrive at a contradiction. Then,

$\Pr\left(\left\{\Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) = \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right) \text{ for some } d \neq d'\right\}\right) = 1.$
Let $\Delta\bar{x}$ be an element of the support of the distribution of Δx. Let $\mathcal{N}_m$ be a sequence of (shrinking) open neighborhoods of $\Delta\bar{x}$. Since $\Delta\bar{x}$ is in the support, $\Pr(\mathcal{N}_m) > 0$. For all m, $\Pr\left(\mathcal{N}_m \cap \left\{\Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) = \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right) \text{ for some } d \neq d'\right\}\right) = \Pr(\mathcal{N}_m) > 0$. Since $\mathcal{N}_m \cap \left\{\Delta x \,\middle|\, \Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) = \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right) \text{ for some } d \neq d'\right\} \neq \emptyset$, for each m there exists an element $\Delta\bar{\bar{x}}_{i,m} \in \mathcal{N}_m$ such that
$\Delta\bar{\bar{\tilde{x}}}_{d_m}'\left(\tilde{\theta}_{d_m} - \tilde{\theta}_{0,d_m}\left(\frac{\breve{\theta}_{d_m}}{\breve{\theta}_{0,d_m}}\right)\right) = \Delta\bar{\bar{\tilde{x}}}_{d_m'}'\left(\tilde{\theta}_{d_m'} - \tilde{\theta}_{0,d_m'}\left(\frac{\breve{\theta}_{d_m'}}{\breve{\theta}_{0,d_m'}}\right)\right)$ for some $d_m \neq d_m'$. Since there are a finite number of pairs of choices d and d', we can find a subsequence m' such that $d_{m'}$ and $d_{m'}'$ are the same for all m'. Since $\Delta\bar{\bar{x}}_{i,m'} \longrightarrow \Delta\bar{x}$, we then have that an arbitrary element of the support, $\Delta\bar{x} \in \left\{\Delta\tilde{x}_d'\left(\tilde{\theta}_d - \tilde{\theta}_{0,d}\left(\frac{\breve{\theta}_d}{\breve{\theta}_{0,d}}\right)\right) = \Delta\tilde{x}_{d'}'\left(\tilde{\theta}_{d'} - \tilde{\theta}_{0,d'}\left(\frac{\breve{\theta}_{d'}}{\breve{\theta}_{0,d'}}\right)\right) \text{ for some } d \neq d'\right\}$. That is, the support of Δx is contained in a collection of $\mathscr{D}(\mathscr{D}+1)/2$ affine hyperplanes. This contradicts Assumption 3(a), so $\Pr\left\{\breve{c} \in \breve{S}(\theta|\Delta\tilde{x})\right\} < 1$. That is, $\Pr\left(\breve{c} \notin \breve{S}(\theta|\Delta\tilde{x})\right) > 0$ and our conclusion follows. □

Proof of Proposition 5: In binary choice, it will be useful note that when $\Pr(y_s = 1|x_s, x_t) = \Pr(y_t = 1|x_s, x_t)$ then $\Pr(y_s = 0|x_s, x_t) = \Pr(y_t = 0|x_s, x_t)$. Take an arbitrary θ. Then, $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}\}, \{\{1\}\}$, or $\{\{0\}, \{1\}\}$. In any of these cases, $H(x_s, x_t, \theta) = 0$, so $H(x_s, x_t, \theta) = 0 \ \forall\theta$ when $\Pr(y_s = 1|x_s, x_t) = \Pr(y_t = 1|x_s, x_t)$.

Another useful finding is that for any θ such that $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}, \{1\}\}$, $H(x_s, x_t, \theta) = \sum_{D \in \overline{\mathbb{D}}(x_s, x_t, \theta)} E[m_D(y_s, y_t) \,|\, x_s, x_t] = [\Pr(y_s = 0|x_s, x_t) - \Pr(y_t = 0|x_s, x_t)] + [\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t)] = 0$.

Take $\theta \in \Theta_0$ and show that $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$ a.s., so that $H(\theta) = H(\theta_0)$. Cases:

(a) $\Delta x_1'\theta > 0$. Then $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{1\}\}$, so $H(x_s, x_t, \theta) = \Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t)$. Since $\theta \in \Theta_0$, $\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) \geq 0$. If $\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) > 0$, then we must have by Proposition 1, $\overline{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{1\}\}$, so $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$. If $\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) = 0$, then as noted above $H(x_s, x_t, \theta) = 0 = H(x_s, x_t, \theta_0)$.

(b) $\Delta x_1'\theta = 0$. Then, $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}, \{1\}\}$, so as noted above $H(x_s, x_t, \theta) = 0$. Since $\theta \in \Theta_0$, we must have $\Pr(y_s = 0|x_s, x_t) - \Pr(y_t = 0|x_s, x_t) \geq 0$ and $\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) \geq 0$. So, $\Pr(y_s = 1|x_s, x_t) = \Pr(y_t = 1|x_s, x_t)$, and $H(x_s, x_t, \theta_0) = 0$. Hence, $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$.

(c) $\Delta x_1'\theta < 0$. Similar to case (a), $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$.

So we have shown that $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$ a.s., and so $H(\theta) = H(\theta_0)$.

Now take $\theta \notin \Theta_0$ and show $H(\theta) < H(\theta_0)$. Let $\mathscr{A} = \{(x_s, x_t) : E[m_D(y_s, y_t) \,|\, x_s, x_t] < 0 \text{ for some } D \in \overline{\mathbb{D}}(x_s, x_t, \theta)\}$. Since $\theta \notin \Theta_0$, $\Pr(\mathscr{A}) > 0$. Consider $(x_s, x_t) \in \mathscr{A}$.

(a) $\Delta x_1'\theta > 0$. But $\Pr(y_s = 1|x_s, x_t) < \Pr(y_t = 1|x_s, x_t)$. Then, $\Pr(y_s = 0|x_s, x_t) > \Pr(y_t = 0|x_s, x_t)$ and $\Delta x_1'\theta_0 < 0$. Hence, $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{1\}\}$ and $\overline{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{0\}\}$, and $H(x_s, x_t, \theta) = \Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) < 0 < \Pr(y_s = 0|x_s, x_t) - \Pr(y_t = 0|x_s, x_t) = H(x_s, x_t, \theta_0)$.

(b) $\Delta x_1'\theta = 0$. Then, $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}, \{1\}\}$, so $H(x_s, x_t, \theta) = 0$. But since $(x_s, x_t) \in \mathscr{A}$, $\Pr(y_s = 1|x_s, x_t) \neq \Pr(y_t = 1|x_s, x_t)$. Suppose $\Pr(y_s = 1|x_s, x_t) > \Pr(y_t = 1|x_s, x_t)$, then $\Delta x_1'\theta_0 > 0$, and $H(x_s, x_t, \theta_0) = \Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t) > 0 = H(x_s, x_t, \theta)$. The argument holds similarly when $\Pr(y_s = 1|x_s, x_t) < \Pr(y_t = 1|x_s, x_t)$. So $H(x_s, x_t, \theta_0) > H(x_s, x_t, \theta)$

(c) $\Delta x_1'\theta < 0$. Arguing similarly to (a), $H(x_s, x_t, \theta_0) > H(x_s, x_t, \theta)$.

So we have shown that $H(x_s, x_t, \theta_0) > H(x_s, x_t, \theta)$ for $(x_s, x_t) \in \mathscr{A}$. For $(x_s, x_t) \notin \mathscr{A}$, it is straightforward to show $H(x_s, x_t, \theta) = H(x_s, x_t, \theta_0)$. Hence, $H(\theta_0) - H(\theta) = E[\mathbf{1}\{(x_s, x_t) \in \mathscr{A}\}(H(x_s, x_t, \theta_0) - H(x_s, x_t, \theta))] > 0$. The result follows. □

Counterexamples

Two examples to show that the criterion $H(\theta)$ defined in section 5 cannot be used for identification for multinomial choice with three or more choices in either sharp- or point- identified cases. Start with the sharp identified case, where only Assumption 1 needs to be satisfied. Then, modify the sharp case to a point identified case with the same conclusion.

In both examples, there will be three choices, $\mathscr{D} + 1 = 3$, and the fixed effects distribution will be a point mass at zero, $\lambda = 0$ with probability one. Covariates will be two-dimensional.

Sharp Identified Example

- Parameter:
$\theta_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. We will also consider the parameter value $\theta = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$.

- Covariate Distribution:

Suppose (x_s, x_t) has the following single point mass distribution.

$$x_{1,t} = x_{2,t} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x_{1,s} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x_{2,s} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \tag{25}$$

Let $\bar{F}_{x_s,x_t}$ denote the c.d.f. of this degenerate distribution, so that $(x_s, x_t) \sim \bar{F}_{x_s,x_t}$

- Disturbance Distribution:

Define the following discrete distribution

$$\Pr((\varepsilon_1, \varepsilon_2) = (a, b)) = \begin{cases} 1/3 & \text{if } (a,b) = (-1,-4) \\ 1/6 & \text{if } (a,b) = (4,0), (0,4), (-4,-4) \\ 1/12 & \text{if } (a,b) = (1,0.5), (-4,-1) \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{F}_\varepsilon$ denote the c.d.f. of this discrete distribution.

Take $\varepsilon_s | x_s, x_t \sim \bar{F}_\varepsilon$ and $\varepsilon_t | x_s, x_t \sim \bar{F}_\varepsilon$. Also, assume ε_s and ε_t are independent conditional on (x_s, x_t). Since $\lambda = 0$, $\varepsilon_s | x_s, x_t \sim \varepsilon_s | x_s, x_t, \lambda$.

- Choice Probabilities:

Given the disturbance distribution (and the covariate values and true parameter value, θ_0), we can compute the conditional choice probabilities.

$$\begin{aligned} \Pr(y_s = 2 \,|\, x_s, x_t, \lambda) &= 1/3, \ \Pr(y_s = 1 \,|\, x_s, x_t, \lambda) \ = \ 1/2, \quad \text{and} \\ \Pr(y_t = 2 \,|\, x_s, x_t, \lambda) &= 1/6, \ \Pr(y_t = 1 \,|\, x_s, x_t, \lambda) \ = \ 1/4, \end{aligned}$$

- Covariate Index Ordering:

Compute the covariate indices, e.g. $\Delta x_2' \theta_0 = (x_{2,s} - x_{2,t})' \theta_0 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}' \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3.$

$$\begin{aligned} \Delta x_2' \theta_0 = 3 \ > \ \Delta x_1' \theta_0 = 2 > 0 \quad &\Longrightarrow \quad \overline{\mathbb{D}}(x_s, x_t, \theta_0) = \{\{2\}, \{1,2\}\}, \text{and} \\ \Delta x_1' \theta = 2 \ > \ \Delta x_2' \theta = 1 > 0 \quad &\Longrightarrow \quad \overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{1\}, \{1,2\}\} \end{aligned}$$

- Criterion:

Let $H_j(x_s, x_t, \theta) = \sum_{|D|=j, D \in \overline{\mathbb{D}}(x_s,x_t,\theta)} E[m_D(y_s, y_t) \,|\, x_s, x_t]$. Then,

$$
\begin{aligned}
H_1(x_s, x_t, \theta_0) &= \Pr(y_s = 2 \,|\, x_s, x_t) - \Pr(y_t = 2 \,|\, x_s, x_t) \;=\; 1/6 \\
H_2(x_s, x_t, \theta_0) &= \Pr(y_s \in \{1, 2\} \,|\, x_s, x_t) - \Pr(y_t \in \{1, 2\} \,|\, x_s, x_t) \;=\; 5/12 \\
&\text{and} \\
H_1(x_s, x_t, \theta) &= \Pr(y_s = 1 \,|\, x_s, x_t) - \Pr(y_t = 1 \,|\, x_s, x_t) \;=\; 1/4 \\
H_2(x_s, x_t, \theta) &= H_2(x_s, x_t, \theta_0) \;=\; 5/12
\end{aligned}
$$

Hence,

$$
\begin{aligned}
H(\theta) - H(\theta_0) &= E[H(x_s, x_t, \theta)] - E[H(x_s, x_t, \theta_0)] \;=\; H(x_s, x_t, \theta) - H(x_s, x_t, \theta_0) \\
&= [H_1(x_s, x_t, \theta) - H_1(x_s, x_t, \theta_0)] + [H_2(x_s, x_t, \theta) - H_2(x_s, x_t, \theta_0)] \\
&= 1/12 + 0 \;=\; 1/12 \;>\; 0
\end{aligned}
$$

Notice that any re-weighting of H_1 and H_2 with positive weights will still yield a positive criterion difference as above.

Point Identified Example

- Parameter:

θ_0 and θ as in Sharp Identified Example.

- Covariate Distribution:

Let Φ_k denote the k-dimensional standard normal distribution. Suppose (x_s, x_t) has a mixture distribution with weights ρ and $1-\rho$, $(x_s, x_t) \sim \rho\Phi_4 + (1-\rho)\bar{F}_{x_s,x_t}$. We will focus on the case where ρ is close to zero.

- Disturbance Distribution:

Let c_x denote the value in the support of $\bar{F}_{x_s,x_t}$ defined in (25). When $(x_s, x_t) = c_x$, take ε_s to have a mixture distribution, $\varepsilon_s|(x_s, x_t) = c_x \;\sim\; \rho\Phi_2 + (1-\rho)\bar{F}_\varepsilon$. Take $\varepsilon_t|(x_s, x_t) = c_x$ to have the same distribution and assume ε_s and ε_t are conditionally independent.
When $(x_s, x_t) \neq c_x$, take ε_s to be multivariate standard normal, $\varepsilon_s|(x_s, x_t) \sim \Phi_2$. As before, take ε_s and ε_t to be conditionally independent and identically distributed.

- Covariate Index Ordering:

When $(x_s, x_t) = c_x$, the covariate index ordering is as in the Sharp Case above, so $\overline{\mathbb{D}}(c_x, \theta_0) = \{\{2\}, \{1, 2\}\}$ and $\overline{\mathbb{D}}(c_x, \theta) = \{\{1\}, \{1, 2\}\}$

- Criterion:

We have used mixture distributions to be sure the support conditions of Assumptions 2 and 3 are satisfied. Since Assumptions 1, 2 and 3 hold, θ_0 is point identified.

$$H(\theta) - H(\theta_0) \geq (1-\rho)^2 1/12 \; - \; 2[1-(1-\rho)^2] \; > \; 0 \quad \text{for } \rho > 0 \text{ small enough}$$

So, even under point identification, $H(\theta)$ is not a useful identification criterion in general.