Moment Inequalities for Multinomial Choice with Fixed Effects

Ariel Pakes
Harvard University and NBER

Jack Porter
University of Wisconsin

September 25, 2021

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. Our main finding shows that the derived conditional moment inequalities yield the sharp identified set for the random utility covariate index, while avoiding the incidental parameter problem. Specializing this result to the binary choice case shows that Manski's (1987) conditional moment inequalities still lead to sharp bounds without restrictions on covariates.

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 covariate index function. 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 (or within group). Under this specification, a novel within-group comparison leads to a set of conditional moment inequalities, which are the basis for our result on sharp partial identification.

Our main finding establishes sharp nonparametric identification of the covariate index function 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.[1]

The semiparametric 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 and in this sense is the discrete choice analog to the familiar within transformation in linear models. As a result, the incidental parameter problem is fully circumvented in our identification results.

The most closely related work is Shi, Shum, and Song (2018), which obtains point identification with a linear covariate index using cyclic monotonicity in the semiparametric multi-

[1]The distribution of unobservables can be expressed as a nonnegative solution to a system of linear equations, so existence is constructive in the sense that a solution can then be determined in a finite number of matrix manipulation steps.

nomial setup. Since our conditional moment inequalities provide sharp bounds, it is not surprising that we are able to show that the Shi, Shum, and Song (2018) conditional moment inequalities are implied by our conditional moment inequalities. It follows that, under the additional conditions on covariates given in Shi, Shum, and Song (2018), our conditional moment inequalities also yield point identification in the linear covariate index case.

When we specialize our setup to the binary choice case with a linear covariate index, we find that our conditional moment inequalities match the weak version of Manski (1987)'s conditional moment inequalities. It follows that these conditional moment inequalities yield sharp bounds even when point identification fails due to either insufficient variation in the covariates or nonlinearities in the covariate index functions. Further, we establish that Manski (1987)'s maximum score criterion can be derived as an aggregation of the conditional moments that make up our inequalities and extended to allow for nonlinear covariate indices. 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.

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 1995). 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).

We apply the findings of this paper to consider whether the price sensitivity of demand for health insurance depends on income in a subsidized health insurance market for low-income consumers. The data come from the Commonwealth Care program in Massachusetts which allowed consumers to choose among competing private plans. We compare the price sensitivity of consumers with incomes one to two times the Federal Poverty Level with those whose income is two to three times the Federal Poverty Level. Implementing the conditional moment inequalities derived here via Andrews and Shi (2013), we obtain a confidence set for the ratio of the price coefficients in the two income groups and reject the hypothesis of no income dependence at the 5% level.

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. Kahn, Ouyang, and Tamer (2019) develop an approach to identification that can be used in both static and dynamic semiparametric multinomial choice models. Chesher, Rosen, and Smolinski (2013) and Chesher and Rosen (2017) obtain sharp identification for nonseparable instrumental variable models that include discrete choice. 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 paper is structured as follows. Section 2 sets up a semiparametric version of 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 (or functions) of interest. In section 4, we address point identification and binary choice. Section 5 implements the conditional moment inequalities in an empirical exercise using data on health insurance choices through the Commonwealth Program exchange in Massachusetts. Section 6 concludes. 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 the Cartesian product of illness category and hospital and t indexing patients, 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,[2] and $2 \leq \mathscr{D} < \infty$.

Given covariates $x_{d,i,t}$ for each choice d associated with observation (i, t), the random utility for choices $d = 0, \ldots, \mathscr{D}$ takes the form

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

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

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 the covariate index function $g_d(\cdot, \theta_0)$, where θ_0 is used to index the functions g_d and is unknown to the researcher. The most commonly assumed form for the index function is linear, e.g. $x'_{d,i,t}\theta_0$. However, the parameter space Θ for θ_0 is unrestricted and need not even be finite-dimensional. That is, θ_0 could index functions in an arbitrary function space, and the index functions are allowed to vary by choice.[3] The additive separability between the covariate index 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. In fact, under the assumptions below, the fixed effect could be absorbed into the disturbance without loss of generality.[4] Normalizations to the model, such as pinning down the scale of coefficients in the linear covariate index case, can be incorporated as restrictions on the space of parameters, covariates, and the dimension of the conditional distribution of unobservables.

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 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, and is useful to extend our results to applications that involve set-valued regressors.[5] To fully specify the choice decision, we adopt a simple rule for resolving ties among maximizing random utility choices. 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 choice with the largest choice index, in this case 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

[3]The index functions could also be allowed to depend on time without any change in the results that follow.

[4]This point will appear again below when we see that the constructed distribution in the proof of sharpness has fixed effects set to zero.

[5]In the set-valued regressors case, the researcher does not know the specific value of some regressors, but does observe a set that contains their values. Pakes and Porter (2014) use the tools developed here to analyze this case. 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).

utility maximizing choices. More generally, any rule for resolving utility maximizing ties that is a non-stochastic function of $argmax_d\, U_{d,i,t}$ will lead to the same form of conditional moment inequalities derived below[6] and so knowledge of such a rule is not needed for their implementation.

The setup thus far is a random utility formulation of multinomial choice except that a choice-specific group fixed effect is included and general covariate indices are allowed. 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 $x_{i,t} = (x'_{0,i,t}, \ldots, x'_{\mathscr{D},i,t})'$, $\lambda_i = (\lambda_{0,i}, \ldots, \lambda_{\mathscr{D},i})$, $\varepsilon_{i,t} = (\varepsilon_{0,i.t}, \ldots, \varepsilon_{\mathscr{D},i,t})$, and $U_{i,t} = (U_{0,i,t}, \ldots, U_{\mathscr{D},i,t})$. 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 $U_{d,i,t} = g_d(x_{d,i,t}, \theta_0) + \lambda_{d,i} + \varepsilon_{d,i,t}$. 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 at $T = 2$. 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 panel data models (Chernozhukov, Fernández-Val, Hahn, and Newey 2013).[7]

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) and for discrete choice in Shi, Shum, and Song (2018). 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

[6] See also footnote 8 for allowable tie-breaking rules.

[7] 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.

case. Importantly, for a given time t, the marginal 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 collection 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,

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

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:

$$\begin{aligned} \mathcal{E}_{d,t} \;=\; & \left\{\varepsilon_t : \varepsilon_{d,t} \;\geq\; \max_{c<d} \left[(g_c(x_{c,t}, \theta_0) - g_d(x_{d,t}, \theta_0)) \;+\; (\lambda_c - \lambda_d) + \varepsilon_{c,t}\right]\right\} \\ & \cap \left\{\varepsilon_t : \varepsilon_{d,t} \;>\; \max_{c>d} \left[(g_c(x_{c,t}, \theta_0) - g_d(x_{d,t}, \theta_0)) \;+\; (\lambda_c - \lambda_d) + \varepsilon_{c,t}\right]\right\}, \end{aligned} \tag{3}$$

where the sets with weak and strict inequalities follow from our rule for resolving utility maximizing ties. 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:

$$d^* \;=\; argmax_c \;\; \left(g_c(x_{c,s}, \theta_0) \;-\; g_c(x_{c,t}, \theta_0)\right). \tag{4}$$

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

$$g_{d^*}(x_{d^*,s}, \theta_0) \;-\; g_{d^*}(x_{d^*,t}, \theta_0) \;\geq\; g_c(x_{c,s}, \theta_0) \;-\; g_c(x_{c,t}, \theta_0), \quad \forall c$$

$$\implies \quad g_c(x_{c,t}, \theta_0) - g_{d^*}(x_{d^*,t}, \theta_0) + (\lambda_c - \lambda_{d^*}) \geq\; g_c(x_{c,s}, \theta_0) - g_{d^*}(x_{d^*,s}, \theta_0) + (\lambda_c - \lambda_{d^*}), \;\; \forall c \tag{5}$$

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. Since the inequality holds regardless of the values of the fixed effects (λ), the fixed effects can be integrated out of the inequality in (6) yielding a corresponding conditional moment inequality below. We also extend the argument behind this inequality to generate additional conditional choice probability comparisons and their related conditional moment inequalities which can then be used for identification of the parameter θ_0.

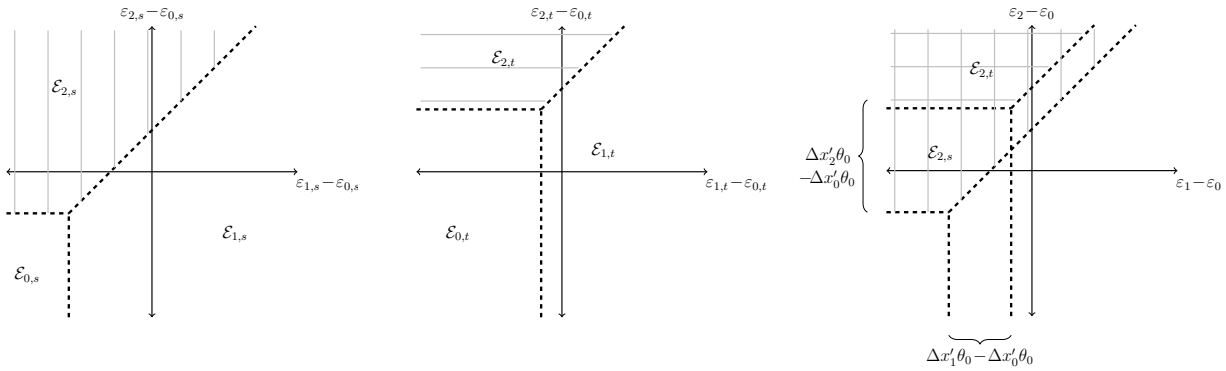
To illustrate the key intuition behind this inequality, consider the case with three choices, a linear covariate index, and $d^* = 2$ implying that $\mathcal{E}_{2,t} \subset \mathcal{E}_{2,s}$. To show the regions $\mathcal{E}_{d,s}$ and $\mathcal{E}_{d,t}$ on two-dimensional graphs, these regions can be re-expressed in terms of $(\varepsilon_{1,s}-\varepsilon_{0,s}, \varepsilon_{2,s}-\varepsilon_{0,s})$ and $(\varepsilon_{1,t}-\varepsilon_{0,t}, \varepsilon_{2,t}-\varepsilon_{0,t})$. In Figure 1a, the time s differenced disturbance space 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 differenced 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 $([x'_{1,s}\theta_0 - x'_{1,t}\theta_0] - [x'_{0,s}\theta_0 - x'_{0,t}\theta_0], [x'_{2,s}\theta_0 - x'_{2,t}\theta_0] - [x'_{0,s}\theta_0 - x'_{0,t}\theta_0])$. The size of these shifts is apparent in Figure 1c; in the case illustrated, $x'_{2,s}\theta_0 - x'_{2,t}\theta_0 \geq x'_{1,s}\theta_0 - x'_{1,t}\theta_0 \geq x'_{0,s}\theta_0 - x'_{0,t}\theta_0$, and 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 complete 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

Figure 1: Disturbance Regions for 3 Choices

$$\Delta x_2'\theta_0 \geq \Delta x_1'\theta_0 \geq \Delta x_0'\theta_0 \quad \Longrightarrow \quad \mathcal{E}_{2,t} \subset \mathcal{E}_{2,s}$$



(a) (b) (c)

index function differences and a set with smaller index function differences. For each such partition generated by differences of the true index functions $g_d(\cdot, \theta_0)$, we will be able to generate corresponding choice probability inequalities.

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. $D \subset \{0, \ldots, \mathscr{D}\}$, and let D^c denote the remaining choices, $D^c = \{0, \ldots, \mathscr{D}\} \backslash D$. If

$$\min_{d \in D} \; g_d(x_{d,s}, \theta) - g_d(x_{d,t}, \theta) \; \geq \; \max_{c \in D^c} \; g_c(x_{c,s}, \theta) - g_c(x_{c,t}, \theta),$$

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,

$$\overline{\mathbb{D}}(x_s, x_t, \theta) = \Bigg\{ D \subset \{0, \ldots, \mathscr{D}\} \;\Bigg|\; D, D^c \neq \varnothing, \quad \min_{d \in D} \; g_d(x_{d,s}, \theta) \; - \; g_d(x_{d,t}, \theta) \; \geq \; \max_{c \in D^c} \; g_c(x_{c,s}, \theta) \; - \; g_c(x_{c,t}, \theta) \Bigg\}. \tag{7}$$

As the notation indicates, this collection of choice sets can be constructed from just the covariates and a parameter value.

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 $C \subset D$.

It is 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: $[g_3(x_{3,s}, \theta) - g_3(x_{3,t}, \theta)] > [g_2(x_{2,s}, \theta) - g_2(x_{2,t}, \theta)] = [g_1(x_{1,s}, \theta) - g_1(x_{1,t}, \theta)] > [g_0(x_{0,s}, \theta) - g_0(x_{0,t}, \theta)]$. 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 $C \subset \{1, \ldots, \mathscr{D}\}$ and a choice d, define

$$\begin{aligned}\mathcal{E}_{d,C,t} \;=\; & \left\{\varepsilon_t : \varepsilon_{d,t} \geq \max_{c \in C: c<d} \left[(g_c(x_{c,t}, \theta_0) - g_d(x_{d,t}, \theta_0)) + (\lambda_c - \lambda_d) + \varepsilon_{c,t}\right]\right\} \\ & \cap \left\{\varepsilon_t : \varepsilon_{d,t} > \max_{c \in C: c>d} \left[(g_c(x_{c,t}, \theta_0) - g_d(x_{d,t}, \theta_0)) + (\lambda_c - \lambda_d) + \varepsilon_{c,t}\right]\right\}\end{aligned}$$

The set $\mathcal{E}_{d,C,t}$ 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 $\mathcal{E}_{d,C,s}$ are defined analogously.

Now take any $D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0)$. For any $d \in D$, $g_d(x_{d,s}, \theta_0) - g_d(x_{d,t}, \theta_0) \geq g_c(x_{c,s}, \theta_0) - g_c(x_{c,t}, \theta_0)$ for all $c \in D^c$. Re-arranging, $g_c(x_{c,t}, \theta_0) - g_d(x_{d,t}, \theta_0) \geq g_c(x_{c,s}, \theta_0) - g_d(x_{d,s}, \theta_0)$ for all $c \in D^c$. It follows that $\mathcal{E}_{d,D^c,t} \subset \mathcal{E}_{d,D^c,s}$ and this

set inclusion is the main step to showing the desired probability inequality:[8]

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

Holding (x_s, x_t, λ) fixed, $\{\varepsilon_s : y_s \in D\} = \cup_{d \in D} \mathcal{E}_{d,D^c,s}$ 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

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

□

The probability inequalities obtained in Proposition 1 can be straightforwardly translated into corresponding moment inequalities, as follows. Let $\mathbb{D} = \{D \,|\, D, D^c \neq \varnothing, D \subset \{0, \ldots, \mathscr{D}\}\}$. For any $D \in \mathbb{D}$, define

$$m_D(y_s, y_t, x_s, x_t, \theta) = \begin{cases} \mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\} & \text{if } D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0) \\ 0 & \text{otherwise} \end{cases}.$$

Then, it follows from Proposition 1 that $E[m_D(y_s, y_t, x_s, x_t, \theta_0) \,|\, x_s, x_t, \lambda] \geq 0, \quad \forall D \in \overline{\mathbb{D}}(x_s, x_t, \theta_0)$. 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):

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

where the inequalities with $D \notin \overline{\mathbb{D}}(x_s, x_t, \theta_0)$ follow immediately from the definition of m_D above. This set of conditional moment inequalities will be the key to the identification

[8] Any rule for resolving random utility ties that preserves this set inclusion, in general, will lead to the same set of conditional moment inequalities.

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}. Let $\mathcal{X}_{s,t}$ denote the support of the joint distribution (x_s, x_t). Define the corresponding identified set as follows:

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

Under Assumption 1 alone, Θ_0 will not generally be a singleton. 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:

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

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,

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

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

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

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 $F_{y_s,y_t,x_s,x_t} \in \mathscr{F}_{ob}$.

□

Theorem 2 is shown by a constructive proof, see Appendix for details. Fix a distribution of observables, $F_{y_s,y_t,x_s,x_t} \in \mathscr{F}_{ob}$. 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 $(\lambda^*, \varepsilon_s^*, \varepsilon_t^*) \,|\, x_s, x_t$ 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 $\mathcal{X}_{s,t}$. We need to choose $(\lambda^*, \varepsilon_s^*, \varepsilon_t^*) \,|\, x_s, x_t$ such that $\Pr(y_s^* = d, y_t^* = d' \,|\, x_s, x_t) = \Pr(y_s = d, y_t = d' \,|\, x_s, x_t)$. Setting $\lambda^* = 0$, $\Pr(y_s^* = d, y_t^* = d' \,|\, x_s, x_t)$ is determined by the behavior of $(\varepsilon_s^*, \varepsilon_t^*) \,|\, x_s, x_t$ on certain "choice-determining" regions of $\mathbb{R}^{2(\mathscr{D}+1)}$. 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. Given the nonparametric flexibility in the disturbance distribution allowed by Assumption 1, 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, if Assumption 1 additionally included the restriction that conditional distributions $(\varepsilon_{i,s}, \varepsilon_{i,s}) \big| x_{i,s}, x_{i,t}, \lambda_i$ are absolutely continuous, Theorem 2 would still hold. The constructed distributions in the proof could be straightforwardly modified, as noted in the Appendix, to show this result. Since random utility ties would be probability zero events under this additional assumption, the model could resolve such ties arbitrarily.

Fifth, the proof outlined above constructs a distribution of unobservables conditional on (x_s, x_t), which shows conditional sharpness. For each value of the covariates, our conditional moment inequalities contain all the information available in the *conditional* distribution of the data. Conditional on any value of the covariates, the parameter region "ruled out" by our conditional moment inequalities must contain the region "ruled out" by any other set of conditional moment inequalities. Since this holds for every covariate value, it follows that the union over all covariate values of ruled out parameter regions for our conditional moment inequalities must contain the union of ruled out parameter regions for any other set of conditional moment inequalities (which is the standard "unconditional" implication of sharpness).

Sixth, 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_0^e and a sharp identified set Θ_S^e. $F^e_{y_s,y_t,x_s,x_t}$ is generated by changing only the marginal distribution of covariates and maintaining the same conditional distribution of unobservables, so Assumption 1 is still satisfied. The conclusion of Theorem 2 follows. Practically this enables the researcher to use the conditional moment inequalities in (10) to consider the identifying power of a particular empirical covariate design and investigate the potential identifying power of alternatives.

4 Additional Remarks

Next, we discuss two topics related to our sharp identification result: point identification and the special case of binary choice.

4.1 Point Identification

In section 3, we showed that the proposed conditional moment inequalities produce a sharp identified set. With a linear covariate index function, point identification is established by imposing further conditions that ensure the identified set reduces to a singleton. Assumption 1 places no restrictions on the covariates. The key additional conditions for point identification ensure sufficient variation in the covariates, in particular an assumption of unboundedness (see Chamberlain 2010).

Shi, Shum, and Song (2018) derived conditional moment inequalities implied by cyclic monotonicity for the multinomial choice model with a linear covariate index function under conditions including Assumption 1 and absolute continuity of the error distribution with respect to Lebesgue measure. Under assumptions on the covariates, they show that their conditional moment inequalities are sufficient for point identification.

It is straightforward to compare the conditional moment inequalities in (10) with a linear covariate index function to the corresponding Shi, Shum, and Song (2018) cyclic monotonicity conditional moment inequalities. We adopt the normalization for choice zero in Shi, Shum, and Song (2018), $x_{0,i,t} = 0$, $\lambda_{0,i} = \varepsilon_{0,i,t} = 0$ so $U_{0,i,t} = 0$ and similarly at time s. From Shi, Shum, and Song (2018) Lemma 3.1, the length 2-cycle conditional moment inequality can be expressed as

$$0 \;\leq\; \sum_{d=1}^{\mathscr{D}} \left[\Pr(y_{i,s} = d \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = d \,|\, x_{i,s}, x_{i,t})\right] \Delta x_d' \theta_0 \tag{13}$$

Now denote a (weak) ordering of covariate index differences as follows:

$$\Delta x'_{(\mathscr{D})^*}\theta_0 \;\geq\; \Delta x'_{(\mathscr{D}-1)^*}\theta_0 \;\geq\; \cdots \;\geq\; \Delta x'_{(0)^*}\theta_0 \tag{14}$$

And suppose that choice zero has the $j{+}1^{th}$ smallest covariate index difference, i.e. $(j)^* \;=\; 0$,

so that $\Delta x'_{(j)^*}\theta_0 = 0$. Then, re-writing the sum in (13),

$$
\begin{aligned}
&\sum_{d=1}^{\mathscr{D}} \Big[\Pr(y_{i,s} = d \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = d \,|\, x_{i,s}, x_{i,t}) \Big] \Delta x'_d \theta_0 \\
&= \sum_{d=j+1}^{\mathscr{D}} \Big[\Pr(y_{i,s} = (d)^* \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = (d)^* \,|\, x_{i,s}, x_{i,t}) \Big] \Delta x'_{(d)^*} \theta_0 \\
&+ \sum_{d=0}^{j-1} \Big[\Pr(y_{i,s} = (d)^* \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = (d)^* \,|\, x_{i,s}, x_{i,t}) \Big] \Delta x'_{(d)^*} \theta_0 \\
&= \sum_{d=j+1}^{\mathscr{D}} \left[\sum_{d'=d}^{\mathscr{D}} \left(\Pr(y_{i,s} = (d')^* \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = (d')^* \,|\, x_{i,s}, x_{i,t}) \right) \right] \left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d-1)^*} \theta_0 \right) \\
&+ \sum_{d=0}^{j-1} \left[\sum_{d'=d}^{\mathscr{D}} \left(\Pr(y_{i,s} = (d')^* \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} = (d')^* \,|\, x_{i,s}, x_{i,t}) \right) \right] \left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d+1)^*} \theta_0 \right) \\
&= \sum_{d=j+1}^{\mathscr{D}} \Big[\Pr(y_{i,s} \in \{(d)^*, \ldots, (\mathscr{D})^*\} \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} \in \{(d)^*, \ldots, (\mathscr{D})^*\} \,|\, x_{i,s}, x_{i,t}) \Big] \left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d-1)^*} \theta_0 \right) \\
&+ \sum_{d=0}^{j-1} \Big[\Pr(y_{i,s} \in \{(0)^*, \ldots, (d)^*\} \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} \in \{(0)^*, \ldots, (d)^*\} \,|\, x_{i,s}, x_{i,t}) \Big] \left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d+1)^*} \theta_0 \right) \\
&= \sum_{d=j+1}^{\mathscr{D}} \Big[\Pr(y_{i,s} \in \{(d)^*, \ldots, (\mathscr{D})^*\} \,|\, x_{i,s}, x_{i,t}) - \Pr(y_{i,t} \in \{(d)^*, \ldots, (\mathscr{D})^*\} \,|\, x_{i,s}, x_{i,t}) \Big] \left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d-1)^*} \theta_0 \right) \\
&+ \sum_{d=0}^{j-1} \Big[\Pr(y_{i,s} \in \{(d+1)^*, ..., (\mathscr{D})^*\} | x_{i,s}, x_{i,t}) - \Pr(y_{i,t} \in \{(d+1)^*, ..., (\mathscr{D})^*\} | x_{i,s}, x_{i,t}) \Big] \left(\Delta x'_{(d+1)^*} \theta_0 - \Delta x'_{(d)^*} \theta_0 \right)
\end{aligned}
\tag{15}
$$

The terms $\left(\Delta x'_{(d)^*} \theta_0 - \Delta x'_{(d-1)^*} \theta_0 \right)$ and $\left(\Delta x'_{(d+1)^*} \theta_0 - \Delta x'_{(d)^*} \theta_0 \right)$ in (15) are nonnegative by the ordering defined in (14). The relationship between the conditional moment inequalities in (10) and the cyclic monotonicity conditional moment inequalities in (13) follows immediately. The inequalities in (10) imply that the probability difference terms in square brackets in (15) are nonnegative, which further implies that (13) holds. We can then conclude that under Assumption 1 and the additional conditions given in Shi, Shum, and Song (2018), the conditional moment inequalities in (10) yield point identification.[9]

The implication illustrated also highlights a point made in the remarks following Theo-

[9] Actually, we find that point identification can be achieved using a subset of the inequalities in (10) that correspond to partitions of the choice set of a fixed size. Fix $\delta \in \{1, \ldots, \mathscr{D}\}$. Then, for point identification, it suffices to consider the subset of conditional moment inequalities in (10) with $|D| = \delta$ or $(\mathscr{D}+1) - \delta$. This collection of conditional moment inequalities is non-nested with the cyclic monotonicity inequalities in (13).

rem 2. The conditional moment inequalities of this paper are actually conditionally sharp. Any given covariate value yields one Shi, Shum, and Song (2018) length 2-cycle conditional moment inequality. Since this inequality is a linear inequality, it rules out a half-space in the parameter space. The same covariate value yields $2^{\mathscr{D}+1} - 2$ conditional moment inequalites from (10) and each of these conditional moment inequalities rules out a cone in the parameter space. Equation (15) shows formally that the union of ruled-out cones must contain the ruled out half-space.

We also find that, in general, the inequalities (13) do not imply (10). Inequality (13) generates the half-space whose boundary is given in (15) (which is a weighted combination of covariate values where the weights are given by the choice probability differences), while the inequalities (10) generate $2^{\mathscr{D}+1}-2$ (cone) regions (each of which is determined by the covariate values and does not depend on the choice probabilities). So in general the boundaries from (15) and (10) will not coincide, i.e. the ruled out half-space given by (15) will be strictly contained in the ruled out region given by (10) for each conditioning value of the covariate.

4.2 Binary Choice

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.[10] Moreover, we show here that Manski's maximum score criterion used for point identification can also be used to obtain sharp partial identification when point identification does not hold. Additionally, while Manski (1987) considers the linear index case, the result shown here allows for parametric or nonparametric covariate index functions as in (1).

In the point identified binary choice case, Manski (1987) proposes an alternative method of estimation, commonly referred to as maximum score estimation. We see below the close connection between the maximum score criterion and the conditional moment inequalities defining the identified set.

The conditional moment inequalities are:

$$E[m_D(y_s, y_t, x_s, x_t, \theta) \,|\, x_s, x_t] \;\geq\; 0, \qquad \forall\; D \in \mathbb{D}, \tag{16}$$

[10] In the binary choice case, Assumption 1(a) is exactly Manski's Assumption 3, and Assumption 1(b) is exactly Manski's Assumption 1(a). So, the sharpness result relaxes Manski's Assumptions 1(b) and 2.

and in the binary choice case,

$$\overline{\mathbb{D}}(x_s, x_t, \theta) = \begin{cases} \{\{1\}\} & \text{for } \Delta g(x_s, x_t, \theta) > 0 \\ \{\{0\}\} & \text{for } \Delta g(x_s, x_t, \theta) < 0 \\ \{\{0\}, \{1\}\} & \text{for } \Delta g(x_s, x_t, \theta) = 0. \end{cases} \tag{17}$$

where

$$\Delta g(x_s, x_t, \theta) = [g_1(x_{1,s}, \theta) - g_1(x_{1,t}, \theta)] - [g_0(x_{0,s}, \theta) - g_0(x_{0,t}, \theta)].$$

The maximum score criterion function is:

$$H(\theta) = E[\text{sgn}(\Delta g(x_s, x_t, \theta))(y_s - y_t)].$$

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

$$\begin{aligned} H(x_s, x_t, \theta) &= \sum_{D \in \mathbb{D}} E[m_D(y_s, y_t, x_s, x_t, \theta) \,|\, x_s, x_t] \\ &= \sum_{D \in \overline{\mathbb{D}}(x_s, x_t, \theta)} E[\mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\} \,|\, x_s, x_t]. \end{aligned} \tag{18}$$

Then,

$$H(\theta) = E[H(x_s, x_t, \theta)]. \tag{19}$$

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. According to Manski (1987) in the binary choice case with a linear covariate index function, under conditions implying point identification, 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 3 *Suppose $\mathscr{D} + 1 = 2$ (binary choice) and Assumption 1 holds. Then,*

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

□

Under the conditions of Theorem 2, Θ_0 is itself sharp, showing that the maximum score criterion will identify the sharp bounds for the covariate index function in the binary choice model.

5 Empirical Example

We implement our conditional moment inequalities in an empirical example that analyzes health insurance choices in the Commonwealth Care (or "CommCare") program in Massachusetts, enacted as part of the state health reforms in 2006. The program provided subsidized health insurance to low-income citizens via an insurance exchange that let consumers choose among competing private plans. We focus on a classic question in demand analysis; does the response of demand to price changes depend on the income of consumers?

CommCare was for citizens whose earnings were less that 300% of the Federal Poverty Level (or the FPL; this was $10,830 in 2010, and increased by the CPI-U annually thereafter). We examine whether the response to price movements differed between two groups of individuals: those with incomes between one and two times the FPL and those whose income was between two and three times the FPL (those with incomes less than the FPL were fully subsidized and hence not included in the analysis). Differences in the price coefficient between these two groups has distributional implications for the welfare generated by this and other programs directed at low income households.

Our model has consumer i at time t choosing among plans d to maximize

$$U_{d,i,t} = -p_{d,i,t}\beta_0 - p_{d,i,t}\mathbf{1}\{I_{i,t} \in [FPL, 2FPL]\}\gamma_0 + \lambda_{d,i} + \varepsilon_{d,i,t}, \tag{20}$$

where individual i's price coefficient is β_0 if that individual's income $I_{i,t}$ is contained in $[FPL, 2FPL]$ and $\beta_0 + \gamma_0$ if $I_{i,t} \in [2FPL, 3FPL]$. The $\lambda_{d,i}$ capture individual (additive) product preferences, and $\varepsilon_{d,i,t}$ captures the remaining unobserved variation in random utility. Our focus is on γ_0, the difference between the price sensitivity of individuals when they are in the higher versus the lower income group. Since the parameters are only identified up to scale, we can only learn about γ_0/β_0. Assuming downward sloping demand ($\beta_0 > 0$) we can normalize $\beta_0 = 1$ so that γ_0 represents the desired ratio.[11]

[11]To explicitly connect the hypotheses on γ_0 with demand effects described, consider the demand effect on choice d of a price decrease in choice d with all other prices staying the same or increasing and income in the low range, let $\Delta_1 = \Pr(y_t = d|p_{d,s}, p_{d,t}, p_{-d,s}, p_{-d,t}, I_s, I_t \in [FPL, 2FPL], \lambda)$ $- \Pr(y_s = d|p_{d,s}, p_{d,t}, p_{-d,s}, p_{-d,t}, I_s, I_t \in [FPL, 2FPL], \lambda)$. Assuming downward sloping demand, $\Delta_1 > 0$. We compare Δ_1 to the change in demand for the same choice under the same price dynamics when in-

We consider regions where four insurers participate in the market during our data period, with each insurer (by rule) offering a single plan. Program rules required each enrollee to make a separate choice; there was no family coverage, and kids were covered in the separate Medicaid program. We analyze plan choices in an annual open enrollment month each year. Individuals are also allowed to choose plans when they change their income group and we treat changes occurring at these times as separate choices for estimation purposes.[12] For more detail on the data and the CommCare program see Shepard (2020), Finkelstein, Hendren, and Shepard (2019), McIntyre, Shepard, and Wagner (2021).

We want to capture choices that are not induced by changes in the individual's choice environment, just by prices, and we require the choice set (plan availability) to be the same four plans in the two periods we compare. We therefore remove comparisons for individuals who changed regions (there are five in the data), or who faced different plan offerings in the comparison periods. We divide the remaining data into cells based on consumers' observed characteristics. The characteristics of a cell are defined by the Cartesian product of; a) pair of years, b) region and c) the income groups in each of the two periods being compared. So the $\lambda_{d,i}$ represent differences in tastes among consumers with the same characteristics. We use all cells as defined above that have more than 20 members.

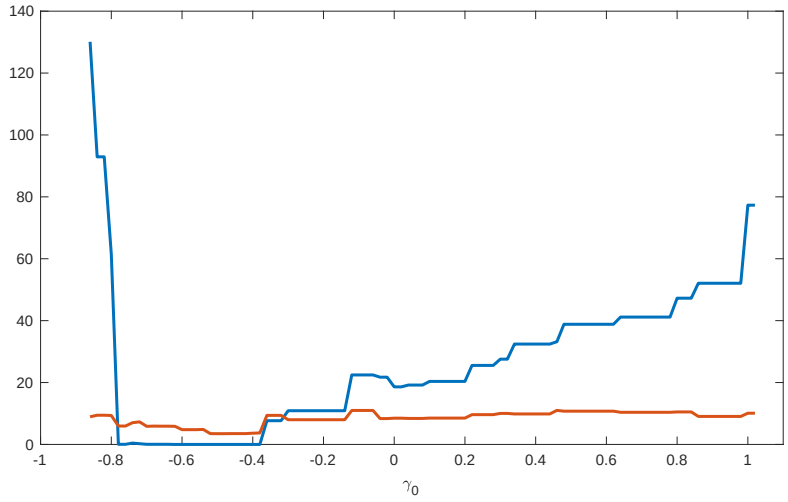
There were large changes in relative prices between 2010 and 2012. During this period, the provider BMC, which had the largest share in 2010 with over a third of the market, increased its average price from below \$50 per member per month to over \$90. By the end of 2012 it was clear that the price increases cost them almost half of their subscribers, and in 2013 they reduced their prices to an average price of just over \$40.[13] We focus on the differential responses to these price changes and use the (s,t) combinations of (2010, 2012) and (2012, 2013). This generates 13,169 pairs of choices and 14 moments corresponding to the four choices.

come increases over time, $\Delta_2 = \Pr(y_t = d|p_{d,s}, p_{d,t}, p_{-d,s}, p_{-d,t}, I_s \in [FPL, 2FPL], I_t \in [2FPL, 3FPL], \lambda)$ $- \Pr(y_s = d|p_{d,s}, p_{d,t}, p_{-d,s}, p_{-d,t}, I_s \in [FPL, 2FPL], I_t \in [2FPL, 3FPL], \lambda)$. Assume that the disturbance distribution is conditionally independent of incomes. Then, under the null $\gamma_0 = 0$, $\Delta_1 = \Delta_2$. Under the alternative that $-\beta_0 < \gamma_0 < 0$, $\Delta_1 > \Delta_2$, that is the increase in the demand is smaller if income increases. And, clearly, both of these conclusions would still hold after integrating the fixed effects out over a given distribution to yield "average" demand effects.

[12]Coverage is heavily regulated, with all cost sharing and covered medical services completely standardized across insurers. The only flexible plan attributes are provider networks. These were largely stable during our sample period with one major exception. Network Health (one of our plans) drops Partners HealthCare (the state's largest medical system) from its hospital network at the start of 2012. To account for this, we treat Network as two different plans, one before and one after 2012, and apply the rules above with that understanding.

[13]There were no major changes in BMC's network or other quality attributes at this time. There was, however, a change in the rules governing the exchange in 2012 which set up an auction like environment to determine the plans that were available to the fully subsidized individuals and likely induced price experimentation.

Figure 2: Criterion Function and Critical Values



The blue line is the negative part criterion function, and the red line is the corresponding 95% bootstrap critical value function.

To use the conditional moment inequalities for inference, there are many recently developed methods (Andrews and Shi 2013, Armstrong 2015, Armstrong and Chan 2016, Chernozhukov, Lee, and Rosen 2013, Chetverikov 2018, and Lee, Song, and Whang 2013). We implement the approach described in Andrews and Shi (2013), which requires a choice of instruments to translate the conditional moments into unconditional moments. We use indicators for year pairs as instruments which yields 28 unconditional moments. Implementing a squared negative part criterion function and bootstrap critical values yields a 95% confidence interval for γ_0, $[-0.79, -0.31]$, indicating that when individuals transit to a higher income group their price sensitivity falls significantly.

More detail is provided in Figure 2. The blue line graphs the sample test statistic. The red line graphs the 95% bootstrap critical values.[14] Using the max criterion function (Armstrong 2014) and using various instruments that aggregate less (yielding more unconditional moments) leads to similarly shaped test statistic graphs (though exact magnitudes depend on the number of unconditional moments). We also considered the Shi, Shum, and Song (2018) length 2-cycle conditional moment inequalities. Using the year pair instruments, as above, yields two linear inequalities. The lower bound information in these moment inequalities is the same as obtained previously. The two linear inequalities are necessarily one-sided inequalities and neither provides upper bound information.[15]

[14] We implement the GMS bootstrap following the recommendations given in Andrews and Shi (2013).

[15] We note that using different instruments we are able to find some upper bound information in the cyclic monotonicity conditional moment inequalities.

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 covariate index function and avoids the incidental parameter problem using only two time periods. These conditional moment inequalities provide sharp bounds without restrictions on the covariates. When $T > 2$, each pair of time periods generates a set of conditional moment inequalities as described above. A sharpness result for this case is left as an open problem for future research.

Our empirical example illustrates how these techniques can be applied to examine differential responses by individuals with different characteristics to a given determinant of choices. This should lead to a better understanding of distributional implications of different policies. 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.

The issue of what is consistent with the "between" variation opens up the question, which we have left for future research, of what information is available on the fixed effects per se. In addition to helping us analyze responses to changes in characteristics, there are cases where knowledge of the fixed effects are of inherent interest and should be analyzable. For example in Ho and Pakes (2014)'s investigation of the impact of capitation on allocation of patients to hospitals, the fixed effects represent the perceived qualities of (the 194) different hospitals for each of (the 106) alternative illness categories. They examine whether the perceptions of the providers from different insurance networks coincide, and are able to rank hospital by their perceived quality for the major illness categories.

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.

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.

Armstrong, T. B. (2014): "Weighted KS statistics for Inference on Conditional Moment Inequalities," *Journal of Econometrics*, 181(2), 92–116.

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

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

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

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

——— (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.

Chesher, A., and A. Rosen (2017): "Generalized Instrumental Variable Models," *Econometrica*, 85(3), 959–989.

Chesher, A., A. Rosen, and K. Smolinski (2013): "An Instrumental Variable Model of Multiple Discrete Choice," *Quantitative Economics*, 4, 157–196.

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

Finkelstein, A., N. Hendren, and M. Shepard (2019): "Subsidizing Health Insurance for Low-income Adults: Evidence from Massachusetts," *American Economic Review*, 109(4), 1530–67.

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*, 104(12), 3841–3884.

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.

McIntyre, A., M. Shepard, and M. Wagner (2021): "Can Automatic Retention Improve Health Insurance Market Outcomes?," Discussion paper, Harvard University Working Paper.

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.

Shepard, M. (2020): "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange," Discussion paper, Harvard University Working Paper.

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: $[g_{(\mathscr{D})}(x_{(\mathscr{D}),s}, \theta) - g_{(\mathscr{D})}(x_{(\mathscr{D}),t}, \theta)] > [g_{(\mathscr{D}-1)}(x_{(\mathscr{D}-1),s}, \theta) - g_{(\mathscr{D}-1)}(x_{(\mathscr{D}-1),t}, \theta)] > \cdots > [g_{(0)}(x_{(0),s}, \theta) - g_{(0)}(x_{(0),t}, \theta)]$ (with some abuse of the 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

$$R_{(\mathscr{D});t} \cup \cdots \cup R_{(d);t} \;\subset\; R_{(\mathscr{D});s} \cup \cdots \cup R_{(d);s}\ , \quad \forall d \in \{1, \ldots, \mathscr{D}\}. \tag{21}$$

Since the sets $R_{(d);s}$ form a partition for $d = 0, \ldots, \mathscr{D}$, the set inclusion (21) implies that

$$R_{(d);s} \cap R_{(d');t} = \emptyset \quad \text{for } d' > d. \tag{22}$$

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. By (22), $q_{(d),(d')\times(d''),(d''')}^* = 0$ if $d < d'$ or $d'' < d'''$, so

$$p_{(d),(d')}^* \;=\; \sum_{\underline{d}=0}^{\mathscr{D}} \sum_{\tilde{d}=0}^{\mathscr{D}} q_{(d),(\underline{d})\times(\tilde{d}),(d')}^* \;=\; \sum_{\underline{d}=0}^{d} \sum_{\tilde{d}=d'}^{\mathscr{D}} q_{(d),(\underline{d})\times(\tilde{d}),(d')}^*.$$

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

$$p_{(d),(d')} \;=\; \sum_{\underset{\sim}{d}=0}^{d}\sum_{\tilde{d}=d'}^{\mathscr{D}} q^*_{(d),(\underset{\sim}{d})\times(\tilde{d}),(d')}, \tag{23}$$

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 $(r_{d,d'}, r_{d'',d'''})$, $\Pr((\varepsilon^*_s, \varepsilon^*_t) = (r_{d,d'}, r_{d'',d'''})|x_s, x_t) = q^*_{d,d'\times d'',d'''}$.[16] 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 $d \geq d'$,

$$0 \;=\; \sum_{\underset{\sim}{d}\leq\tilde{d}} \left(q^*_{(\tilde{d}),(\underset{\sim}{d})\times(d),(d')} - q^*_{(d),(d')\times(\tilde{d}),(\underset{\sim}{d})} \right). \tag{24}$$

In addition to equations (23) and (24), 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 (23) can be restated as $p = Q_s q^*$, and let Q_p be the matrix with elements in $\{-1, 0, 1\}$ such that equation (24) 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 $q^*_{(d),(d')\times(d),(d')}$ with $d \geq d'$ appear in (A) but not (B). Hence, $z_{(d),(d')} \geq 0$

[16] A continuous distribution for $(\varepsilon^*_s, \varepsilon^*_t)|x_s, x_t$ could be obtained by defining the density on each $R_{d,d'} \neq \emptyset$ to be a constant chosen so that $\Pr((\varepsilon^*_s, \varepsilon^*_t) \in R_{d,d'}|x_s, x_t) = q^*_{d,d'\times d'',d'''}$.

for $d \geq d'$. 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

$$\sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\tilde{d}),(\underset{\sim}{d})} \;\geq\; \sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\underset{\sim}{d}),(\tilde{d})} \quad \text{for } d = 1, \ldots, \mathscr{D} \tag{25}$$

Now, for constants a_d,

$$\begin{aligned} z'p &= \sum_{d=0}^{\mathscr{D}} z_{(d),(d)} p_{(d),(d)} + \sum_{d>d'} \left(z_{(d),(d')} p_{(d),(d')} + z_{(d'),(d)} p_{(d'),(d)} \right) \\ &\geq \sum_{d>d'} \left(z_{(d),(d')} p_{(d),(d')} + z_{(d'),(d)} p_{(d'),(d)} \right) \\ &= \sum_{d=1}^{\mathscr{D}} a_d \underbrace{\left[\sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\tilde{d}),(\underset{\sim}{d})} - \sum_{\tilde{d}=d}^{\mathscr{D}} \sum_{\underset{\sim}{d}=0}^{d-1} p_{(\underset{\sim}{d}),(\tilde{d})} \right]}_{\geq 0 \text{ by (25)}} \\ &\quad + \sum_{s=1}^{\mathscr{D}} \sum_{d=0}^{\mathscr{D}-s} \left\{ \left(z_{(d+s),(d)} - \left[\sum_{d'=d+1}^{d+s} a_{d'} \right] \right) p_{(d+s),(d)} + \left(z_{(d),(d+s)} + \left[\sum_{d'=d+1}^{d+s} a_{d'} \right] \right) p_{(d),(d+s)} \right\} \end{aligned}$$

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 $d + 1 \leq d'$, let $\underline{a}_{d',d+1} = \min_{\tilde{d} \geq d', \underset{\sim}{d} \leq d} \left\{ w_{(\tilde{d}),(d')} - w_{(d),(\underset{\sim}{d})} \right\}$ and $\underline{a}_{d'} = \max\{0, \underline{a}_{d',d'}\}$. Then, for $d > d'$, $z_{(d),(d')} \geq \bar{a}_{d,d'+1}$. And, for $d < d'$, $-z_{(d),(d')} \leq \underline{a}_{d',d+1}$, where we have imposed (25) through the definition of $\underline{a}_{d'}$. 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 $\underline{a}_d \leq a_d \leq \bar{a}_d$ for $d = 1, \ldots, \mathscr{D}$, and $\underline{a}_{d+s,d+1} \leq \sum_{d'=d+1}^{d+s} a_{d'} \leq \bar{a}_{d+s,d+1}$ 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, by another version of Farkas' Lemma, it is sufficient

to show that

$$\sum_{d=1}^{\mathscr{D}} b_d \underline{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} b_{d',d} \underline{a}_{d',d} \leq \sum_{d=1}^{\mathscr{D}} c_d \bar{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} c_{d',d} \bar{a}_{d',d} \tag{26}$$

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

$$b_d + \sum_{1 \leq d' \leq d \leq d'' \leq \mathscr{D}} b_{d'',d'} \;=\; c_d + \sum_{1 \leq d' \leq d \leq d'' \leq \mathscr{D}} c_{d'',d'}. \tag{27}$$

Before showing (26), it will be useful to first note a fact about the coefficients $b_\cdot$ and $c_\cdot$. Let $h_d = \min\{b_d, c_d\}$ and $h_{d',d} = \min\{b_{d',d}, c_{d',d}\}$ for all d', d. Equation (26) can be re-stated as

$$\begin{aligned}
&\sum_{d=1}^{\mathscr{D}} (b_d - h_d) \underline{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} (b_{d',d} - h_{d',d}) \underline{a}_{d',d} + \left[\sum_{d=1}^{\mathscr{D}} h_d \underline{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} h_{d',d} \underline{a}_{d',d} \right] \\
&\quad \leq \;\sum_{d=1}^{\mathscr{D}} (c_d - h_d) \bar{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} (c_{d',d} - h_{d',d}) \bar{a}_{d',d} + \left[\sum_{d=1}^{\mathscr{D}} h_d \bar{a}_d \;+\; \sum_{d=1}^{\mathscr{D}-1} \sum_{d'=d+1}^{\mathscr{D}} h_{d',d} \bar{a}_{d',d} \right].
\end{aligned}$$

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

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

So it will suffice to show

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

Or, more simply, in proving (26), 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 (26) follows.

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

where the second equality follows from differences of the equations given in (27).

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 Proposition 3: In binary choice, it will be useful to note that when $\Pr(y_s = 1|x_s, x_t) = \Pr(y_t = 1|x_s, x_t)$ then $\Pr(y_s = 0|x_s, x_t) = \Pr(y_t = 0|x_s, x_t)$. In this case, 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 $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}, \{1\}\}$, $H(x_s, x_t, \theta) = \sum_{D \in \overline{\mathbb{D}}(x_s, x_t, \theta)} E[\mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\} \,|\, x_s, x_t] = [\Pr(y_s = 0|x_s, x_t) - \Pr(y_t = 0|x_s, x_t)] + [\Pr(y_s = 1|x_s, x_t) - \Pr(y_t = 1|x_s, x_t)] = 0$.

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) $\Delta g(x_s, x_t, \theta) > 0$. 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) $\Delta g(x_s, x_t, \theta) = 0$. 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) $\Delta g(x_s, x_t, \theta) < 0$. 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 $\mathcal{A} = \{(x_s, x_t) : E[m_D(y_s, y_t, x_s, x_t, \theta) \,|\, x_s, x_t] < 0 \text{ for some } D \in \mathbb{D}\} = \{(x_s, x_t) : E[\mathbf{1}\{y_s \in D\} - \mathbf{1}\{y_t \in D\} \,|\, x_s, x_t] < 0 \text{ for some } D \in \overline{\mathbb{D}}(x_s, x_t, \theta)\}$. Since $\theta \notin \Theta_0$, $\Pr(\mathcal{A}) > 0$.[17] Consider $(x_s, x_t) \in \mathcal{A}$.

(a) $\Delta g(x_s, x_t, \theta) > 0$. 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 $\Delta g(x_s, x_t, \theta_0) < 0$. 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) $\Delta g(x_s, x_t, \theta) = 0$. Then, $\overline{\mathbb{D}}(x_s, x_t, \theta) = \{\{0\}, \{1\}\}$, so $H(x_s, x_t, \theta) = 0$. But since $(x_s, x_t) \in \mathcal{A}$, $\Pr(y_s = 1|x_s, x_t) \neq \Pr(y_t = 1|x_s, x_t)$. Suppose $\Pr(y_s = 1|x_s, x_t) > \Pr(y_t = 1|x_s, x_t)$, then $\Delta g(x_s, x_t, \theta_0) > 0$, 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) $\Delta g(x_s, x_t, \theta) < 0$. Arguing similarly to (a), $H(x_s, x_t, \theta_0) > H(x_s, x_t, \theta)$.

[17] Suppose $\mathcal{A}$ is measurable or contains a measurable set of positive measure.

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