Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity*

Xiaoxia Shi
University of Wisconsin-Madison

Matthew Shum
Caltech

Wei Song
University of Wisconsin-Madison

May 11, 2017

Abstract

This paper proposes a new semi-parametric identification and estimation approach to multinomial choice models in a panel data setting with individual fixed effects. Our approach is based on *cyclic monotonicity*, which is a defining convex-analytic feature of the random utility framework underlying multinomial choice models. From the cyclic monotonicity property, we derive identifying inequalities without requiring any shape restrictions for the distribution of the random utility shocks. These inequalities point identify model parameters under straightforward assumptions on the covariates. We propose a consistent estimator based on these inequalities.

Keywords: Cyclic Monotonicity, Multinomial Choice, Panel Data, Fixed Effects, Convex Analysis.

*Emails: xshi@ssc.wisc.edu, mshum@caltech.edu, wsong22@wisc.edu. We thank Khai Chiong, Federico Echenique, Bruce E. Hansen, Jack R. Porter, and seminar audiences at Brown, Michigan State, Ohio State, Chicago Booth, Johns Hopkins, Northwestern, NYU, UC Riverside, UNC, the 2016 Seattle-Vancouver Econometrics Conference and the 2015 Xiamen/WISE Econometrics Conference in Honor of Takeshi Amemiya, the 2016 Interaction Conference at Northwestern, and the 2016 Conference on Nonstandard Problems at Duke for useful comments. Alejandro Robinson-Cortes, Pengfei Sui and Jun Zhang provided excellent research assistance. Xiaoxia Shi acknowledges the financial support of the Wisconsin Alumni Research Foundation via the Graduate School Fall Competition Grant.

1 Introduction

Consider a panel multinomial choice problem where agent i chooses from $K+1$ options (labeled $k = 0, \ldots, K$). Choosing option k in period t gives the agent indirect utility

$$\beta' X_{it}^k + A_i^k + \epsilon_{it}^k, \tag{1.1}$$

where X_{it}^k is a d_x-dimensional vector of observable covariates that has support $\mathcal{X}$, β is the vector of weights for the covariates in the agent's utility, $\mathbf{A}_i = (A_i^0, \ldots, A_i^K)'$ are agent-specific fixed effects, and ϵ_{it}^k are unobservable utility shocks the distribution of which is not specified. The agent chooses the option that gives her the highest utility:

$$Y_{it}^k = 1\{\beta' X_{it}^k + A_i^k + \epsilon_{it}^k \geq \beta' X_{it}^{k'} + A_i^{k'} + \epsilon_{it}^{k'};\ \forall k'\}, \tag{1.2}$$

where Y_{it}^k denotes the multinomial choice indicator. Let the data be identically and independently distributed (i.i.d.) across i. As is standard in semiparametric settings, we normalize $\|\beta\| = 1$, $X_{it}^0 = \mathbf{0}_{d_x}$ and $A_i^0 = \epsilon_{it}^0 = 0$. We do not impose any location normalization on ϵ_{it}^k or A_i^k, and as a result, it is without loss of generality to assume that X_{it}^k does not contain a constant.

In this paper, we propose a new semi-parametric approach to the identification and estimation of β. We exploit the notion of *cyclic monotonicity*, which is an appropriate generalization of "monotonicity" to multivariate (i.e. vector-valued) functions. The notion has not been used as a tool for the identification and estimation of semi-parametric multinomial choice models, although the cyclic monotonicity between consumption and price in a representative consumer basket has been used in econometrics as early as Browning (1989) for testing rational expectation hypotheses.

In cross-sectional multinomial models, it is easy to show that there is a cyclic monotonicity relationship between the conditional choice probability and the utility index vector under independence between the unobservable shocks and the utility indices. We apply that to the panel model given above, find a way to integrate out the fixed effects, and obtain a collection of conditional moment inequalities which, conveniently, are linear in β. Then we show that these moment inequalities point identify β under either of two sets of primitive verifiable conditions. We finally propose a consistent estimator for β, the computation of which requires only convex optimization and thus is not sensitive to starting values of the optimization routine.

This paper is most closely related to several contemporaneous papers. Pakes and Porter (2015) propose a different approach to construct moment inequalities for the panel data multinomial

choice model, based on ranking the options according to their conditional choice probabilities. By comparison, we compare the entire vector of choice probabilities for all options across time periods. Khan, Ouyang, and Tamer (2016) propose an approach to point identification in a dynamic panel setting. Some of their identification strategies are similar to ours, but our estimators are rather different.

Our paper builds upon the existing literature on semi-parametric panel binary choice models. Manski (1987) proposed the maximum score approach for identification and estimation. Honoré and Kyriazidou (2000) use a maximum score-type estimator for a dynamic panel binary choice model. Abrevaya (2000) proposes a general class of rank-correlation estimators, which is a smoothed version of Manski's (1987) estimator when applied to the panel binary choice models. Honoré and Lewbel (2002) generalize the special regressor approach of Lewbel (1998, 2000) to the panel data setting.

Semi-parametric identification and estimation of multinomial choice models have been considered in cross-sectional settings (i.e., models without individual fixed effect). Manski (1975) and Fox (2007) base identification on the assumption of a *rank-order property* that the ranking of $\beta' X_i^k$ across k is the same as that of $E[Y_i^k|X_i]$ across k; this is an IIA-like property that allows utility comparisons among all the options in the choice set to be decomposed into pairwise comparisons among these options. To ensure this rank-order property, Manski assumes that the error terms are i.i.d. across k, while Fox relaxes the i.i.d. assumption to exchangeability. Exchangeability (or the rank-order property) is not used in our approach. Lewbel (2000) considers identification using a special regressor. In addition, Powell and Ruud (2008) and Ahn, Ichimura, Powell, and Ruud (2015) consider an approach based on matching individuals with equal conditional choice probabilities, which requires that the rank of a certain matrix formed from the data to be deficient by exactly 1. This approach does not obviously extend to the panel data setting with fixed effects.

The existing literatures on cross-sectional binary choice models and on the semi-parametric estimation of single or multiple index models (which include discrete choice models as examples) is voluminous and less relevant for us, and thus is not reviewed here for brevity.[1]

The paper proceeds as follows. In section 2, we introduce the notion of cyclic monotonicity and relate it to panel multinomial choice models with fixed effects. Subsequently, in Section 3, we present the moment inequalities emerging from cyclic monotonicity, and give assumptions under which these

[1] An exhaustive survey is provided in Horowitz (2009), chapters 2 and 3.

inequalities suffice to point identify the parameters of interest. This section also contains some numerical illustrations. Section 4 presents an estimator, shows its consistency, and evaluates its performance using Monte Carlo experiments. In Section 5, we discuss the closely related aggregate panel multinomial choice model, which is a workhorse model for demand modeling in empirical IO. This section also contains an illustrative empirical application using aggregate supermarket scanner data. Section 6 concludes.

2 Preliminaries

In this section, we describe the concept of cyclic monotonicity and its connection to multinomial choice models. We begin by providing the definition of cyclic monotonicity.

Definition 1 (Cyclic Monotonicity)**.** *Consider a function $f : \mathcal{U} \to R^K$ where $\mathcal{U} \subseteq R^K$, and a length M-cycle of points in R^K: $u_1, u_2, \ldots, u_M, u_1$. The function f is cyclic monotone with respect to the cycle $u_1, u_2, \ldots, u_M, u_1$ if* [2]

$$\sum_{m=1}^{M} (u_m - u_{m+1})' f(u_m) \geq 0, \tag{2.1}$$

where $u_{M+1} = u_1$. The function f is cyclic monotone on $\mathcal{U}$ if it is cyclic monotone with respect to all possible cycles of all lengths on its domain.

Cyclic monotonicity is defined for mappings from $R^K \to R^K$, which generalizes the usual monotonicity for real-valued functions. We make use of the following basic result which relates cyclic monotonicity to convex functions:

Proposition 1 (Cyclic monotonicity and Convexity)**.** *Consider a differentiable function $F : \mathcal{U} \to R$ for an open convex set $\mathcal{U} \subseteq R^K$. If F is convex on $\mathcal{U}$, then the gradient of F (denoted $\nabla F(u) := \partial F(u)/\partial u$) is cyclic monotone on $\mathcal{U}$.*

The proof for Proposition 1 is available from standard sources (e.g, Rockafellar (1970, Ch. 24), Villani (2003, Sct. 2.3)). Consider a univariate and differentiable convex function; obviously, its slope must be monotonically nondecreasing. The above result states that cyclic monotonicity is the appropriate extension of this feature to multivariate convex functions.

[2] Technically, this defines the property of being "cyclic monotonically increasing," but for notational simplicity and without loss of generality, we use "cyclic monotone" for "cyclic monotonically increasing."

Now we connect the above discussion to the multinomial choice model. We start with a generic random utility model for multinomial choices without specifying the random utility function or the data structure in detail. Suppose that an agent is choosing from $K+1$ choices $0, 1, \ldots, K$. The utility that she derives from choice k is partitioned into two additive parts: $U^k + \epsilon^k$, where U^k denotes the systematic component of the latent utility, while ϵ^k denotes the random shocks, idiosyncratic across agents and choice occasions. She chooses choice k^* if $U^{k^*} + \epsilon^{k^*} \geq \max_{k=0,\ldots,K} U^k + \epsilon^k$. Let $Y^k = 1$ if she chooses choice k and 0 otherwise. As is standard, we normalize $U^0 = \epsilon^0 = 0$.

Let u^k denote a generic realization of U^k. Also let $\mathbf{U} = (U^1, \ldots, U^K)'$, $\mathbf{u} = (u^1, \ldots, u^K)'$, and $\boldsymbol{\epsilon} = (\epsilon^1, \ldots, \epsilon^K)'$. We introduce the "social surplus function" (McFadden (1978, 1981)), which is the expected utility obtained from the choice problem:

$$\mathcal{W}(\mathbf{u}) = E\left\{ \max_{k=0,\ldots,K} [U^k + \epsilon^k] | \mathbf{U} = \mathbf{u} \right\}. \tag{2.2}$$

The following lemma shows that this function is convex and differentiable, that its gradient corresponds to the choice probability function, and finally that the choice probability function is cyclic monotone. The first three parts of the lemma are already known in the literature (eg. McFadden (1981)), and the last part is immediately implied by the previous parts and Proposition 1. Nonetheless, we give a self-contained proof in Appendix A for easy reference for the reader.

Lemma 2.1 (Gradient)**.** *Suppose that* $\mathbf{U}$ *is independent of* $\boldsymbol{\epsilon}$ *and that the distribution of* $\boldsymbol{\epsilon}$ *is absolutely continuous with respect to the Lebesgue measure. Then*

(a) $\mathcal{W}(\cdot)$ *is convex on* R^K*,*

(b) $\mathcal{W}(\cdot)$ *is differentiable on* R^K*,*

(c) $\mathbf{p}(\mathbf{u}) = \nabla \mathcal{W}(\mathbf{u})$*, where* $\mathbf{p}(\mathbf{u}) = E[\mathbf{Y}|\mathbf{U} = \mathbf{u}]$ *and* $\mathbf{Y} = (Y^1, \ldots, Y^K)'$*, and*

(d) $\mathbf{p}(\mathbf{u})$ *is cyclic monotone on* R^K*.*

The cyclic monotonicity of the choice probability can be used to form identifying restrictions for the structural parameters in a variety of settings. In this paper, we focus on the linear panel data model with fixed effects, composed of equations (1.1) and (1.2).

3 Panel Data Multinomial Choice Models with Fixed Effects

We focus on a short panel data setting where there are only two time periods. An extension to multiple time periods is given in Section 5. Let $\mathbf{U}$, $\boldsymbol{\epsilon}$, and $\mathbf{Y}$ be indexed by both i (individual)

and t (time period). Thus they are now $\mathbf{U}_{it} \equiv (U_{it}^1, \ldots, U_{it}^K)'$, $\boldsymbol{\epsilon}_{it} \equiv (\epsilon_{it}^1, \ldots, \epsilon_{it}^K)'$, and $\mathbf{Y}_{it} \equiv (Y_{it}^1, \ldots, Y_{it}^K)'$. Let there be an observable d_x dimensional covariate X_{it}^k for each choice k, and let U_{it}^k be a linear index of X_{it}^k plus an unobservable individual effect A_i^k:

$$U_{it}^k = \beta' X_{it}^k + A_i^k, \tag{3.1}$$

where β is a d_x-dimensional unknown parameter. Let $\mathbf{X}_{it} = (X_{it}^1, \ldots, X_{it}^K)$ and $\mathbf{A}_i = (A_i^1, \ldots, A_i^K)'$. Note that $\mathbf{X}_{it}$ is a $d_x \times K$ matrix. In short panels, the challenge in this model is the identification of β while allowing correlation between the covariates and the individual effects. We tackle this problem using the cyclic monotonicity of the choice probability, as we explain next.

3.1 Identifying Inequalities

We derive our identification inequalities under the following assumption.

Assumption 3.1. (a) $\boldsymbol{\epsilon}_{i1}$ *and* $\boldsymbol{\epsilon}_{i2}$ *are identically distributed conditional on* $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$*:*

$$(\boldsymbol{\epsilon}_{i1} \sim \boldsymbol{\epsilon}_{i2}) | (\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2})$$

(b) *the conditional distribution of* $\boldsymbol{\epsilon}_{it}$ *given* $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$ *is absolutely continuous with respect to the Lebesgue measure for* $t = 1, 2$ *everywhere on the support of* $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$*.*

Remark. (i) Part (a) of the assumption is the multinomial version of the the group homogeneity assumption of Manski (1987), and is also imposed in Pakes and Porter (2015). It allows us to form identification inequalities based on the comparison of choices made by the same individual over different time periods, and by doing this to eliminate the fixed effect. This assumption rules out dynamic panel models where X_{it}^k may include lagged values of $(Y_{it}^k)_{k=1}^K$.But it allows $\boldsymbol{\epsilon}_{it}$ to be correlated with the covariates, and arbitrary dependence between $\boldsymbol{\epsilon}_{it}$ and the fixed effects.

(ii) The assumption imposes no restriction on the dependence amongst the errors. The errors across choices in a given period can have arbitrarily dependent, and the errors across time periods, although assumed to have identical marginal distributions, can have arbitrary dependence. ■

To begin, we let $\boldsymbol{\eta}$ denote a K dimensional vector with the kth element being η^k, and define

$$\mathbf{p}(\boldsymbol{\eta}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}) := \Big(\Pr[\epsilon_{i1}^k + \eta^k \geq \epsilon_{i1}^{k'} + \eta^{k'} | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2, \mathbf{A}_i = \mathbf{a}] \Big)_{k=1,\ldots,K}. \tag{3.2}$$

Assumption 3.1(a) implies that

$$\mathbf{p}(\boldsymbol{\eta}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}) = \Big(\Pr[\epsilon_{i2}^k + \eta^k \geq \epsilon_{i2}^{k'} + \eta^{k'} | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2, \mathbf{A}_i = \mathbf{a}] \Big)_{k=1,\ldots,K}. \tag{3.3}$$

Assumption 3.1(b) implies that $\mathbf{p}(\boldsymbol{\eta}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a})$ is cyclic monotone in $\boldsymbol{\eta}$ for all possible values of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}$. Using the cyclic monotonicity with respect to length-2 cycles, we obtain, for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}$:

$$(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)'[\mathbf{p}(\boldsymbol{\eta}_1, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}) - \mathbf{p}(\boldsymbol{\eta}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a})] \geq 0. \tag{3.4}$$

Now we let $\eta_1 = \mathbf{X}_{i1}'\beta + A_i$ and $\eta_2 = \mathbf{X}_{i2}'\beta + A_i$. Note that for $t = 1, 2$, by the definition of $\mathbf{p}(\eta, \mathbf{x}_1, \mathbf{x}_2, a)$, we have

$$\mathbf{p}(\mathbf{X}_{it}'\beta + \mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{A}_i) = E[\mathbf{Y}_{it} | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{A}_i]. \tag{3.5}$$

Combining (3.4) and (3.5), we have

$$(E[\mathbf{Y}_{i1}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{A}_i] - E[\mathbf{Y}_{i2}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{A}_i])(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta) \geq 0 \text{ everywhere.} \tag{3.6}$$

Note that the fixed effect differences out within the second parenthetical term on the left hand-side. Take the conditional expectation given $\mathbf{X}_{i1}, \mathbf{X}_{i2}$ of both sides, and we get,

$$(E[\mathbf{Y}_{i1}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] - E[\mathbf{Y}_{i2}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}])(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta) \geq 0 \text{ everywhere.} \tag{3.7}$$

These inequality restrictions involve only identified/observed quantities and the unknown parameter β, and thus can be used to set identify β in the absence of further assumptions, and to point identify β with additional assumptions as discussed below. Note that under binary choice ($K = 1$), both terms on the LHS of (3.7) become scalars, so that these inequalities reduce to the rank correlation result in Manski (1987, Lemma 1).

Hence the foregoing derivations have proved the following lemma:

Lemma 3.1. *Under Assumption* 3.1*,*

$$(E[\mathbf{Y}_{i1}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] - E[\mathbf{Y}_{i2}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}])(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta) \geq 0 \textit{ pointwise.}$$

The extension in Section 5 discusses how longer cycles can be used when more time periods are available in the dataset. The next subsection presents conditions under which length-2 cycles are enough to produce point identification.

3.2 Point Identification of Model Parameters

To study the identificaiton information contained in the inequalties in (3.7) contain, we rewrite them as

$$E[\Delta \mathbf{Y}_i' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \Delta \mathbf{X}_i' \beta \geq 0 \tag{3.8}$$

where $\Delta Z_i = Z_{i2} - Z_{i1}$.

Define $\mathbf{g} \equiv (\Delta \mathbf{X}_i E[\Delta \mathbf{Y}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}])$. For identification, we will want to place restrictions on the support of the vector $\mathbf{g}$, which we define as:

$$\mathcal{G} = \text{supp}(\mathbf{g}) = \text{supp}(\Delta \mathbf{X}_i E[\Delta \mathbf{Y}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}]). \tag{3.9}$$

We would like to find conditions on model primitives ($\mathbf{X}_{it}$, $\mathbf{A}_{it}$ and $\boldsymbol{\epsilon}_{it}$) that guarantee that the support of the vectors $\mathbf{g}$ is rich enough to ensure point identification.

First, we impose regularity conditions on the unobservables:

Assumption 3.2. (a) *The conditional support of* $(\boldsymbol{\epsilon}_{it} | \mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}) = R^K$ *for a set of values of* $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$ *with positive measure.*

(b) *The conditional distribution of* $(\boldsymbol{\epsilon}_{it}, A_i)$ *given* $(\mathbf{X}_{i1}, \mathbf{X}_{i2}) = (\mathbf{x}_1, \mathbf{x}_2)$ *is uniformly continuous in* $(\mathbf{x}_1, \mathbf{x}_2)$*. That is,*

$$\lim_{(\mathbf{x}_1, \mathbf{x}_2) \to (\mathbf{x}_1^0, \mathbf{x}_2^0)} \sup_{\mathbf{e}, \mathbf{a} \in R^K} |F_{\boldsymbol{\epsilon}_{it}, \mathbf{A}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}}(\mathbf{e}, \mathbf{a} | \mathbf{x}_1, \mathbf{x}_2) - F_{\boldsymbol{\epsilon}_{it}, \mathbf{A}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}}(\mathbf{e}, \mathbf{a} | \mathbf{x}_1^0, \mathbf{x}_2^0)| = 0.$$

The role of Assumption 3.2(a) becomes clear when we describe the covariate conditions below. Assumption 3.2(b) is a sufficient condition for the continuity of the function $E[\Delta \mathbf{Y}_i | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2]$. The latter ensures that the violation of the inequality $E[\Delta \mathbf{Y}_i' | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2] \Delta \mathbf{x}' b \geq 0$ for a point $(\mathbf{x}_1, \mathbf{x}_2)$ on the support of $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ implies that the inequality $E[\Delta \mathbf{Y}_i' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \Delta \mathbf{X}_i' b \geq 0$ is violated with positive probability.

We also need a condition on the observable $\Delta \mathbf{X}_i$. In general this is not straightforward. Note that the vectors $\mathbf{g}$ are equal to

$$\Delta \mathbf{X}_i E[\Delta \mathbf{Y}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}] = \sum_{k=1}^{K} \Delta X_i^k E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \tag{3.10}$$

In general, it is difficult to formulate conditions on the RHS of the previous equation because the RHS is a weighted sum of ΔX_i^k where the weight is the conditional choice probability, which is

not a primitive quantity. We proceed by considering two approaches to reduce the RHS to a single term.

There are two types of events conditional on which we can reduce the summation to a single term:

1. For a given k, let $\Delta\mathbf{X}_i^{-k} = (\Delta X_i^1, \ldots, \Delta X_i^{k-1}, \Delta X_i^{k+1}, \ldots, \Delta X_i^K)$. Conditional on the event $\Delta\mathbf{X}_i^{-k} = 0$ (that is, individual i's covariates are constant across both periods, for all choices except the k-th choice), we have

$$\Delta\mathbf{X}_i E[\Delta\mathbf{Y}_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}] = \Delta X_i^k E[\Delta Y_i^k|\mathbf{X}_{i1}, \mathbf{X}_{i2}].$$

   Consequently, $\text{supp}(\Delta X_i^k E[\Delta Y_i^k|\mathbf{X}_{i1}, \mathbf{X}_{i2}]) = \text{supp}(\Delta X_i^k \text{sign}(E[\Delta Y_i^k|\mathbf{X}_{i1}, \mathbf{X}_{i2}]))$. Assumption 3.2(a) ensures that $\Pr(E[\Delta Y_i^k|\mathbf{X}_{i1}, \mathbf{X}_{i2}] = 0|\Delta X_i^{-k} = 0) = 0$, which implies that $\text{sign}(E[\Delta Y_i^k|\mathbf{X}_{i1}, \mathbf{X}_{i2}]) \in \{-1, 1\}$ with probability one conditional on $\Delta\mathbf{X}_i^{-k} = 0$. Thus, it is sufficient to assume a rich support for ΔX_i^k and $-\Delta X_i^k$ conditional on $\Delta\mathbf{X}_i^k$. We thus define

$$G_I \equiv \cup_k \text{supp}(\pm\Delta X_i^k|\Delta\mathbf{X}_i^{-k} = 0). \tag{3.11}$$

2. Conditional on the event $\Delta X_i^k = \Delta X_i^1$ for all k (that is, individual i's covariates are identical across all choices and only vary across time periods), we have

$$\Delta\mathbf{X}_i E[\Delta\mathbf{Y}_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}] = \Delta X_i^1 E[-\Delta Y_i^0|\mathbf{X}_{i1}, \mathbf{X}_{i2}],$$

   where $\Delta Y_i^0 = -\sum_{i=1}^K \Delta Y_i^k$. Consequently, we have

$$\text{supp}(\Delta X_i^1 E[\Delta Y_i^0|\mathbf{X}_{i1}, \mathbf{X}_{i2}]) = \text{supp}(\Delta X_i^1 \text{sign}(E[\Delta Y_i^0|\mathbf{X}_{i1}, \mathbf{X}_{i2}])). \tag{3.12}$$

   Similar to above, we define

$$G_{II} \equiv \text{supp}(\pm\Delta X_i^1|\Delta X_i^k = \Delta X_i^1 \ \forall k). \tag{3.13}$$

In what follows, our identification condition will be imposed on the set

$$G \equiv G_I \cup G_{II}. \tag{3.14}$$

Two assumptions on G are considered, which differ in the types of covariates that they accommodate. Each assumption is sufficient by itself. We consider each case in turn.

Assumption 3.3. *The set G contains an open R^{d_x} ball around the origin.*

The gist of this assumption is that, beginning from the origin and moving in any direction, we will reach a point in G. This assumption essentially requires all covariates to be continuous, but allows them to be bounded.[3]

Our second sufficient condition allows discrete covariates generally, but requires one regressor with large support.[4] Let g_{-1} denote g with the first element removed, and define $G_{-1} = \{g_{-1} : \exists g_1 \in R \; s.t. (g_1, g_{-1}')' \in G\}$. Let $G_1(g_{-1}) = \{g_1 \in R : (g_1, g_{-1}')' \in G\}$. For $j = 2, \ldots, d_x$, we define g_{-j}, G_{-j}, and $G_j(g_{-j})$ analogously.

Assumption 3.4. *For some $j^* \in \{1, 2, \ldots, d_x\}$:*

(a) *$G_{j^*}(g_{-j^*}) = \mathbb{R}$ for all g_{-j^*} in a subset $G^0_{-j^*}$ of G_{-j^*},*

(b) *$G^0_{-j^*}$ is symmetric about the origin, and is not contained in a proper linear subspace of $R^{d_x - 1}$,*

(c) *the j^*th element of β, denoted by β_{j^*}, is nonzero.*

The identification result is stated using the following criterion function:

$$Q(b) = E\Big| \min(0, E[\Delta \mathbf{Y}_i' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \Delta \mathbf{X}_i' b) \Big|. \tag{3.15}$$

We will return to this criterion function below in considering estimation.

Theorem 3.1. *Under Assumptions* 3.1, 3.2*, and either* 3.3 *or* 3.4*, we have $Q(\beta) = 0$, and $Q(b) > 0$ for all $b \neq \beta$ such that $b \in R^{d_x}$ and $\|b\| = 1$.*

3.3 Examples

Next we consider several examples, which show that verifying Assumption 3.3 or 3.4 can be straightforward. For all the examples, we consider the trinary choice ($K = 2$) case with two covariates ($d_x = 2$).

[3] In the binary case, this set of conditions reduces to conditions similar to those in Höderlein and White (2012).

[4] In the binary choice case, this set of conditions reduces to conditions similar to those in Manski(1987). Allowing for discrete covariates generally requires the presence of an unbounded covariate to achieve uniform identification (Chamberlain (2010)). This impossibility result is implied by our necessity result (Theorem B.1 in Appendix B) which shows that uniform point identification is impossible if all regressors are bounded and at least one of them is finite-valued (e.g. the time dummy). When no regressor is finite-valued, boundedness does not preclude point identification, as shown in Assumption 3.3.

Example 1. $\text{supp}((X_{it}^k)_{t=1,2;k=1,2}) = [0,1]^8$. Then $\text{supp}((\Delta X_i^k)_{k=1,2}) = [-1,1]^4$. Then, $G_I = \text{supp}(\Delta X_i^2 | \Delta X_i^1 = 0) = [-1,1]^2$. Obviously, $[-1,1]^2$ contains an open neighborhood of the origin; thus, Assumption 3.3 is satisfied.

Example 2. Suppose that the covariates do not vary across k: $X_{it}^k = X_{it}$ for $k = 1,2$, and $\text{supp}((X_{it})_{t=1,2}) = [0,1]^4$. Thus, $G_{II} = \text{supp}(\Delta X_i) = [-1,1]^2$ which satisfies Assumption 3.3.

Example 3. Suppose that the covariates take continuous values for alternative 1 and discrete values for alternative 2, as an example of which $\text{supp}((X_{it}^1)_{t=1,2}) = [0,1]^4$, $\text{supp}((X_{it}^2)_{t=1,2}) = \{0,1\}^4$, and the joint support is the Cartesian product. Then, $\text{supp}(\Delta X_i^1 | \Delta X_i^2 = \mathbf{0}) = [-1,1]^2$. Thus, Assumption 3.3 is satisfied.

Example 4. Suppose that the first covariate is a time dummy: $X_{1,it}^k = t$ for all k, t, and the second covariate has unbounded support: $\text{supp}((X_{2,it}^k)_{t=1,2;k=1,2}) = (c,\infty)^4$ for some $c \in R$. Then,

$$\text{supp}\left(\Delta X_i^1 | \Delta X_i^1 = \Delta X_i^2\right) = \{1\} \times \mathbb{R}.$$

Hence, $G \supseteq G_{II} = \{-1,1\} \times \mathbb{R}$. Let $j^* = 2$ (for j^* as defined in Assumption 3.4), and let $G_{-2}^0 = \{-1,1\}$. Assumption 3.4(b) obviously holds. Assumption 3.4(a) also holds because $G_2(-1) = G_2(1) = \mathbb{R}$. Assumption 3.4(c) holds as long as $\beta_2 \neq 0$.

3.4 Remarks: Cross-Sectional Model

In this paper we have focused on identification and estimation of *panel* multinomial choice models. Here we briefly remark on the use of the CM inequalities for estimation in cross-sectional multinomial choice models, which is natural and can be compared to the large number of existing estimators for these models. In the cross-sectional model, the individual-specific effects disappear, leading to the choice model

$$Y_i^k = 1\{\beta' X_i^k + \epsilon_i^k \geq \beta' X_i^{k'} + \epsilon_i^{k'} \text{ for all } k' = 0, \ldots, K\}.$$

Hence, to apply the CM inequalities, the only dimension upon which we can difference is across individuals. Under the assumptions that the vector of utility shocks $\boldsymbol{\epsilon}_i$ is (i) i.i.d. across individuals and (ii) independent of the covariates $\mathbf{X}$, the 2-cycle CM inequality yields that, for all pairs (i,j),

$$(E[\mathbf{Y}_i|\mathbf{X}_i] - E[\mathbf{Y}_j|\mathbf{X}_j]) \cdot (\mathbf{X}_i - \mathbf{X}_j)' \beta \geq 0.$$

In particular, for the binary choice case ($k \in \{0,1\}$), this reduces to

$$\left(E[Y_i^1|\mathbf{X}_i] - E[Y_j^1|\mathbf{X}_j]\right) \cdot (\mathbf{X}_i - \mathbf{X}_j)'\beta \geq 0$$

which is the estimating equation underlying the maximum score (Manski (1975)) and maximum rank correlation (Han (1987)) estimators for the binary choice model.

4 Estimation and Consistency

Since the identification in this paper is based on inequalities rather than equalities, standard estimation and inference methods do not apply. Nevertheless, we propose a computationally easy consistent estimator for β, based on Theorem 3.1.

In the asymptotic analysis, we consider the case of a short panel; that is, the number of time period T is fixed and the number of agents $n \to \infty$. In particular, we consider $T = 2$. Based on the panel data set, suppose that there is a uniformly consistent estimator $\hat{\mathbf{p}}_t(\mathbf{x}_1, \mathbf{x}_2)$ for $E(\mathbf{Y}_{it}|\mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2)$ for $t = 1, 2$. Then we can estimate the model parameters using a sample version of the criterion function given in equation (3.15). Specifically, we obtain a consistent estimator of β as $\widehat{\beta} = \overline{\beta}/\|\overline{\beta}\|$, where

$$\overline{\beta} = \arg \min_{b \in R^{d_x} : \max_{j=1,\ldots,d_x} |b_j| = 1} Q_n(b), \quad \text{and} \tag{4.1}$$

$$Q_n(b) = n^{-1} \sum_{i=1}^{n} \left[(b' \Delta \mathbf{X}_i)(\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2})) \right]_- , \tag{4.2}$$

where $[z]_- = |\min(0, z)|$, and $\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2}) = \hat{\mathbf{p}}_2(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \hat{\mathbf{p}}_1(\mathbf{X}_{i1}, \mathbf{X}_{i2})$. The estimator is easy to compute because $Q_n(b)$ is a convex function and the constraint set of the minimization problem is the union of $2d_x$ convex sets. If one knows the sign of a parameter, say $\beta_1 > 0$, one can simplify the estimator even further by using the constraint set $\{b \in R^{d_x} : b_1 = 1\}$ instead. [5] [6]

The following theorem shows the consistency of $\widehat{\beta}$.

[5] An alternative candidate for $\widehat{\beta}$ is $\arg\min_{b \in R^{d_x} : \|b\| = 1} Q_n(b)$. However, obtaining this estimator requires minimizing a convex function on a non-convex set, which is computationally less attractive.

[6] Instead of forming the criterion function using a nonparametric estimator of $\mathbf{p}(\cdot, \cdot)$, one could also use weight functions to turn the conditional inequalities into unconditional inequalities, as done in Khan and Tamer (2009) and Andrews and Shi (2013). We investigate this option in the Monte Carlo experiment and report the results in Appendix C.

Assumption 4.1. (a) $\sup_{\mathbf{x}_1,\mathbf{x}_2\in\text{supp}(\mathbf{X}_{i1},\mathbf{X}_{i2})} \|\Delta\hat{\mathbf{p}}(\mathbf{x}_1,\mathbf{x}_2)-\Delta\mathbf{p}(\mathbf{x}_1,\mathbf{x}_2)\| \to_p 0$ *as* $n\to\infty$ *for* $t=1,2$*, where* $\Delta\mathbf{p}(\mathbf{x}_1,\mathbf{x}_2)=E[\mathbf{Y}_{i2}-\mathbf{Y}_{i1}|\mathbf{X}_{i1}=\mathbf{x}_1,\mathbf{X}_{i2}=\mathbf{x}_2]$ *for* $t=1,2$*, and*

(b) $\max_{t=1,2} E[\|\mathbf{X}_{it}\|]<\infty$.

Theorem 4.1 (Consistency)**.** *Under Assumptions* 3.1*,* 3.2*,* 4.1*, and either* 3.3 *or* 3.4*:*

$$\widehat{\beta}\to_p \beta \quad \text{as} \quad n\to\infty.$$

The consistency result in Theorem 4.1 relies on a uniformly consistent estimator of the change of the conditional choice probability $\Delta\mathbf{p}(\mathbf{x}_1,\mathbf{x}_2)$. Such estimators are abundant in the nonparametric regression literature; see for example, Cheng (1984) for the k-nearest neighbor estimator, Chapter 2 of Li and Racine (2006) for kernel regression estimators, and Hirano, Imbens, and Ridder (2009) for a sieve logit estimator.

Remark: Partial identification. Here we have focused on point identification of the model parameters utilizing the cyclic monotonicity inequalities. An alternative would be to consider the case when the parameters are partially identified. In that case, confidence intervals for β can be constructed using the methods proposed for general conditional moment inequalities (see, for example, Andrews and Shi (2013) and Chernozhukov, Lee and Rosen (2013)). These methods are partial-identification robust, and thus can be applied when our point identification assumptions hold or do not hold. Moreover, since our moment inequalities, based on cyclic monotonicity, are linear in the model parameters, we can also utilize more specialized methods for models with nonsingleton, convex identified sets (Bontemps, Magnac, and Maurin (2012); Freyberger and Horowitz (2015)). These methods may involve easier computation than the general methods. ■

4.1 Monte Carlo Simulation

Consider a trinary choice example and a two-period panel. Let X_{it}^k be a three-dimensional covariate vector: $X_{it}^k=(X_{j,it}^k)_{j=1,2,3}$. Let $(X_{j,it}^k)_{j=1,2,3;k=1,2;t=1,2}$ be independent uniform random variables in $[0,1]$.[7] Let $A_i^k=(\omega_i^k+X_{1,i2}-X_{1,i1})/4$ for $k=1,2$, where ω_i^k is uniform in $[0,1]$, independent across k and independent of other model primitives. Consider the true parameter value $\beta=(1,0.5,0.5)$, and use the scale normalization $\beta_1=1$.

[7] Assumption 3.3 is satisfied because all the X variables are supported on the unite interval and they can vary freely from each other. Thus point identification holds under this design.

We consider two specifications. The first specification is a multinomial logit model. In the second specification, ϵ_{it}^k for $k = 1, 2$ is a difference of two independent Cauchy$(x_0 = 0, \gamma = 2)$ variates.

In addition to our CM estimator, we also implement Chamberlain's (1980) conditional likelihood estimator for comparison. The conditional likelihood method is consistent and $n^{-1/2}$-normal for the logit specification, but it may not be consistent in the Cauchy specification.[8] For both estimators, we report bias, standard deviation (SD) and the root mean-squared error (rMSE). To implement our estimator, we use the Nadaraya Watson estimator with product kernel to estimate $\mathbf{p}(\cdot, \cdot)$ with bandwidth selected via leave-one-out cross-validation. We consider four sample sizes 250, 500, 1000, and 2000, and use 5000 Monte Carlo repetitions.

The results are reported in Tables 1 and 2. We only report the performance of $\widehat{\beta}_2$ because that of $\widehat{\beta}_3$ is nearly the same due to the symmetric design of the experiment. Under the Logit design (Table 1), the conditional likelihood estimator not surprisingly has smaller bias and smaller standard deviation. Yet our CM estimator is close in performance with conditional likelihood. Under the Cauchy design, the conditional likelihood estimator displays larger bias and standard deviation, and the bias shrinks very slowly with the sample size. This may reflect the inconsistency of the conditional likelihood estimator in this set up. On the other hand, the CM estimator has a smaller bias and standard deviation, both decreasing significantly as the sample size increases. Overall, our CM estimator has more robust performance in non-logit setup and leads to only modest efficiency loss in the logit setups in the range of sample sizes that we consider.[9]

[8] In Appendix C, we report an instrumental function variation of our estimator, where the conditional moment inequalities are approximated by unconditional moment inequalities generated by multiplying the moment function to indicators of hypercubes on the space of the conditioning variables in the spirit of Khan and Tamer (2009) and Andrews and Shi (2013), instead of estimated nonparametrically. This variation of our estimator is more difficult to compute and has less desirable Monte Carlo performance.

[9] As the sample size gets larger, the discrepancy between the standard deviation of the CM estimator and the conditional likelihood estimator may grow because the latter is $n^{-1/2}$-consistent while the former likely converges slower.

Table 1: Monte Carlo Performance of Estimators of β_2 (Logit Design, $\beta_{0,2} = 0.5$)

n	BIAS	SD	rMSE	25% quantile	median	75% quantile
CM Estimator						
250	-0.0622	0.1385	0.1519	0.3435	0.4302	0.5242
500	-0.0484	0.0977	0.1090	0.3854	0.4477	0.5141
1000	-0.0328	0.0701	0.0774	0.4192	0.4647	0.5133
2000	-0.0257	0.0488	0.0552	0.4402	0.4726	0.5069
Conditional Likelihood Estimator						
250	0.0064	0.1283	0.1284	0.4192	0.4992	0.5862
500	0.0022	0.0888	0.0889	0.4419	0.5009	0.5581
1000	0.0016	0.0621	0.0621	0.4592	0.5004	0.5430
2000	-0.0001	0.0439	0.0439	0.4700	0.4994	0.5287

Table 2: Monte Carlo Performance of Estimators of β_2 (Cauchy Design, $\beta_{0,2} = 0.5$)

n	BIAS	SD	rMSE	25% quantile	median	75% quantile
CM Estimator						
250	-0.1164	0.2156	0.2450	0.2393	0.3761	0.5226
500	-0.0698	0.1714	0.1851	0.3124	0.4237	0.5379
1000	-0.0392	0.1291	0.1349	0.3701	0.4587	0.5454
2000	-0.0151	0.0953	0.0965	0.4209	0.4809	0.5462
Conditional Likelihood Estimator						
250	0.1791	0.5985	0.6247	0.4118	0.6014	0.8467
500	0.1304	0.2512	0.2830	0.4613	0.6018	0.7607
1000	0.1166	0.1642	0.2013	0.5038	0.6045	0.7182
2000	0.1110	0.1142	0.1593	0.5327	0.6042	0.6809

5 Longer Panels ($T > 2$)

We have thus far focused on two-period panel data sets for ease of exposition. Our method naturally generalizes to longer panel data sets as well. Suppose that there are T time periods. Then one can

use all cycles with length $L \leq T$ to form the moment inequalities. To begin, consider $t_1, t_2, \ldots, t_L \in \{1, 2, \ldots, T\}$, where the points do not need to be all distinct. Assuming the multi-period analogue of Assumption 3.1, we can use derivation similar to that in Section 3.1 to obtain

$$\sum_{m=1}^{L} \beta'(\mathbf{X}_{it_m} - \mathbf{X}_{it_{m+1}}) E[\mathbf{Y}_{it_m} | \mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L}] \geq 0. \tag{5.1}$$

To form an estimator, we consider an estimator $\hat{\mathbf{p}}(\mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L})$ of $E[\mathbf{Y}_{it_m} | \mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L}]$. Let the sample criterion function be

$$Q_n(b) = n^{-1} \sum_{i=1}^{n} \sum_{t_1, \ldots, t_L \in \{1, \ldots, T\}} \left[\sum_{m=1}^{L} b'(\mathbf{X}_{it_m} - \mathbf{X}_{it_{m+1}}) \hat{\mathbf{p}}(\mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L}) \right]_- \tag{5.2}$$

The estimator of β, $\widehat{\beta}$ is defined based on $Q_n(b)$ in the same way as in Section 4.

If $L = T$, the estimator just defined utilizes all available inequalities implied by cyclic monotonicity. However, in practice there are disadvantages of using long cycles because (1) the estimator $\hat{\mathbf{p}}(\mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L})$ can be noisy when $t_1, \ldots, t_L$ contains many distinct values, and (2) it is computationally more demanding to exhaust and aggregate all cycles of longer length if T is moderately large. Thus, in the empirical application below, we only use the length-2 cycles, that is, $L = 2$.

For identification, it might be possible to obtain weaker conditions for point identification when longer cycles are used, but we were not able to come up with a clean set of conditions for that. For estimation, inequalities from longer cycles provide additional restriction on the parameter and thus could potentially improve efficiency. We investigate the gain in a Monte Carlo experiment next.

Another interesting question is whether (5.1) with $L = T$ exhaust all the information in the random utility model and leads to the sharp identified set. We believe this is unlikely because the CM inequalities only derive from the convexity of the social surplus function, and do not use other properties of the random utility model. For instance, in random utility models, the choice probability of one alternative should not increase when its own utility index stays constant while the utility indices of the other alternatives weakly increase, which is not captured in the CM inequalities.[10] However, these properties are not straightforward to use in the panel data setting and do not lead to simple linear (in parameters) moment conditions.

[10] These other properties are studied in Koning and Ridder (2003).

5.1 Monte Carlo Results Using Longer Cycles

Here we use a 3-period extension of the Cauchy design presented in the previous section. All the specification details are the same (including the fact that A_i^k depends only on $X_{1,i2}^k - X_{1,i1}^k$), except that one additional period of data is generated. In Table 3, we report the performance of our moment inequality estimators for β_2 using length-2 cycles, and using both length-2 and 3 cycles (all cycles). As we can see, the performance of the estimator is nearly identical whether or not the length-3 cycles are used. In practice, one can try using length-2 cycles only first and then add length-3 cycles to see if the results change. If not, there should be no reason to consider longer cycles since longer cycles involve higher computational cost.

Table 3: Moment Inequality Estimator of β_2 ($T = 3$, Cauchy Design, $\beta_{0,2} = 0.5$)

n	BIAS	SD	rMSE	25% quantile	median	75% quantile
Based on Length-2 Cycles Only						
250	-0.1413	0.1393	0.1984	0.2631	0.3565	0.4506
500	-0.0989	0.1069	0.1457	0.3283	0.3997	0.4716
1000	-0.0693	0.0814	0.1069	0.3755	0.4300	0.4837
2000	-0.0467	0.0616	0.0773	0.4120	0.4529	0.4936
Based on All cycles						
250	-0.1436	0.1391	0.1999	0.2613	0.3553	0.4465
500	-0.1006	0.1069	0.1467	0.3254	0.3973	0.4683
1000	-0.0702	0.0817	0.1077	0.3736	0.4291	0.4837
2000	-0.0478	0.0618	0.0782	0.4108	0.4514	0.4920

6 Related model: Aggregate Panel Multinomial Choice Model

Up to this point, we have focused on the setting when the researcher has individual-level panel data on multinomial choice. In this section, we discuss an important and simpler related model: the panel multinomial choice model estimated using *aggregate* data for which we are able to derive

some inference results. Such models are often encountered in empirical industrial organization.[11] In this setting, the researcher observes the aggregated choice probabilities (or *market shares*) for the consumer population in a number of regions and across a number of time periods. Correspondingly, the covariates are also only observed at region/time level for each choice option. To be precise, we observe $(\boldsymbol{S}_{ct}, \mathbf{X}_{ct} = (X_{ct}^{1,\prime}, \ldots, X_{ct}^{K,\prime})')_{c=1,t=1}^{n,T}$ which denote, respectively, the region/time-level choice probabilities and covariates. Only a "short" panel is required, as our approach works with as few as two periods. Thus, to get the idea across with the simplest notation possible, we focus on the case where $T = 2$.

We model the individual choice $\mathbf{Y}_{ict} = (Y_{ict}^1, \ldots, Y_{ict}^K)'$ as

$$Y_{ict}^k = 1\{\beta' X_{ct}^k + A_{ic}^k + \epsilon_{ict}^k \geq \beta' X_{ct}^{k'} + A_{ic}^{k'} + \epsilon_{ict}^{k'} \ \forall k' = 0, \ldots, K\}, \tag{6.1}$$

where X_{ct}^0, A_{ic}^0, and ϵ_{ict}^0 are normalized to zero, $\mathbf{A}_{ic} = (A_{ic}^0, \ldots, A_{ic}^K)'$ is the choice-specific individual fixed effect, and $\boldsymbol{\epsilon}_{ict} = (\epsilon_{ict}^1, \ldots, \epsilon_{ict}^K)'$ is the vector of idiosyncratic shocks. Correspondingly, the vector of choice probabilities $\boldsymbol{S}_{ct} = (S_{ct}^1, \ldots, S_{ct}^K)'$ is obtained as the fraction of I_{ct} agents in region i and time t who chose option k, i.e. $\boldsymbol{S}_{ct} = I_{ct}^{-1} \sum_{i=1}^{I_{ct}} \mathbf{Y}_{ict}$.

Make the market-by-market version of Assumption 3.1:

Assumption 6.1. (a) *The error terms are identically distributed* $(\boldsymbol{\epsilon}_{ic1} \sim \boldsymbol{\epsilon}_{ic2})$ *conditional on market and individual identity. Let market identity be captured by a random element* η_c*; then this condition can be written as* $(\boldsymbol{\epsilon}_{ic1} \sim \boldsymbol{\epsilon}_{ic2})|\eta_c, A_{ic}$ *and*

(b) *the conditional c.d.f. of* $\boldsymbol{\epsilon}_{ict}$ *given* $\mathbf{A}_{ic}, \eta_c$ *is absolutely continuous with respect to the Lebesgue measure, everywhere in* $\mathbf{A}_{ic}, \eta_c$*.*

Then arguments similar to those for Lemma 3.1 imply the following lemma.

Lemma 6.1. *Under Assumption* 6.1*, we have*

$$E(\Delta \mathbf{Y}_{ic}'|\eta_c)(\Delta \mathbf{X}_c'\beta) \geq 0, \ a.s. \tag{6.2}$$

We no longer need to perform the nonparametric estimation of conditional choice probabilities because $E(\mathbf{Y}_{ict}|\eta_c)$ can be estimated uniform consistently by $\boldsymbol{S}_{ct}$.[12]

[11] See, for instance, Berry, Levinsohn, and Pakes (1995) and Berry and Haile (2014).

[12] If $\inf_{c,t} I_{ct}$ grows fast enough with $n \times T$, this estimator is uniformly consistent, i.e. $\sup_c \sup_t \|\boldsymbol{S}_{ct} - E(\mathbf{Y}_{ict}|\eta_c)\| \to_p 0$. Section 3.2 of Freyberger's (2013) arguments (using Bernstein's Inequality) imply that the above convergence holds if $\log(n \times T)/\min_{c,t} I_{ct} \to 0$.

Now, we can construct a consistent estimator of β. The estimator is defined as

$$\widehat{\beta} = \overline{\beta}/\|\overline{\beta}\|, \quad \text{where} \tag{6.3}$$

$$\overline{\beta} = \arg \min_{b \in R^{d_x} : \max_{j=1,\ldots,J} |b_j|=1} Q_n(b) = n^{-1} \sum_{c=1}^{n} \left[(b' \Delta \mathbf{X}_c)(\Delta \boldsymbol{S}_c) \right]_- . \tag{6.4}$$

This estimator is consistent by similar arguments as those for Theorem 4.1. Estimators using longer cycles when $T > 2$ can be constructed as in the previous section.

6.1 Convergence Rate of $\widehat{\beta}$ in the Aggregate Case

In the aggregate case, I_{ct} is typically large relative to n. As a result, it is often reasonable to assume that I_{ct} increases fast as $n \to \infty$, and S_{ct} converges to the limiting choice probability $E[\mathbf{Y}_{ict}|\eta_c]$ fast enough that the difference between S_{ct} and $E[\mathbf{Y}_{ict}|\eta_c]$ has negligible impact on the convergence of $\widehat{\beta}$. Under such assumptions, we can derive a $n^{-1/2}$ convergence rate for $\widehat{\beta}$.

The derivation involves differentiating the criterion function with respect to b, which is easier to explain on a convex parameter space rater than the unit circle that we have been using as the normalized parameter space. Thus, for ease of exposition, we switch to the normalization $\beta_1 = 1$ in this section. The parameter space is hence $\{(1, \tilde{b}')' : \tilde{b} \in R^{d_x - 1}\}$. Let $\tilde{\beta}$ denote β with the first coordinate removed. We make the following assumptions.

Assumption 6.2. (a) $(\mathbf{S}_{ct}, \mathbf{X}_{ct})_{t=1}^2$ *is i.i.d. (independent and identically distributed) across* c*, and* $E(\|\text{vec}(X_{ct})\|^2) < \infty$.

(b) $\max_{t=1,2} E[\|\mathbf{S}_{ct} - E[\mathbf{Y}_{ict}|\eta_c]\|^2] = O(n^{-1})$.

(c) $\widehat{\beta} \to_p \beta$.

Let $\mathbf{W}_c = (\Delta \mathbf{X}_c) E[\Delta \mathbf{Y}_{ic}|\eta_c]$. *Let* $\mathbf{W}_c^1$ *denote the first coordiante of* $\mathbf{W}_c$*, and let* $\tilde{\mathbf{W}}_c$ *denote* $\mathbf{W}_c$ *with the first coordinate removed.*

(d) $\Pr(b'\mathbf{W}_c = 0) = 0$ *for all* b *such that* $b_1 = 1$ *and* $\|b - \beta\| \leq c_1$ *for a constant* c_1.

(e) *With probability one, the conditional c.d.f.* $F_{\mathbf{W}_c^1|\tilde{\mathbf{W}}_c}(\cdot|\tilde{\mathbf{W}}_c)$ *of* $\mathbf{W}_c^1$ *given* $\tilde{\mathbf{W}}_c$ *is continuous on* $[-\tilde{\beta}'\tilde{\mathbf{W}}_c, \infty)$*, continuously differentiable on* $(-\tilde{\beta}'\tilde{\mathbf{W}}_c, \infty)$ *with the derivative* $f_{\mathbf{W}_c^1|\tilde{\mathbf{W}}_c}(\cdot|\tilde{\mathbf{W}}_c)$ *that is bounded by a constant* C.

(f) *The smallest eigenvalue of*

$$E[\tilde{\mathbf{W}}_c \tilde{\mathbf{W}}_c' f_{\mathbf{W}_c^1|\tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c'\tilde{\beta} - \tau \mathbf{W}_c'(b - \beta)|\tilde{\mathbf{W}}_c) 1(\mathbf{W}_c'(b - \beta) < 0)]$$

is bounded below by $c_2 > 0$ *for all* $\tau \in (0,1)$ *and all* b *such that* $b_1 = 1$ *and* $\|b - \beta\| \leq c_1$.

For establishing the rate result, we follow the general methods of Kim and Pollard (1990) and Sherman (1993), which are useful for dealing with the noise component due to sample averaging in the criterion function (6.4). This is the only source of noise we need to consider, as Assumption 6.2(b) ensures that the noise from using the observed market shares $\mathbf{S}_{ct}$ to estimate the conditional expectations $E[\mathbf{Y}_{ct}|\eta_c]$ is negligible.[13]

Parts (d)-(f) of this assumption require further explanation. We need to establish a quadratic lower bound for the limiting criterion function in a neighborhood of the true value β. We do so via deriving the first and the second order directional derivatives of the limiting criterion function in such a neighborhood. Parts (d)-(e) are used to guarantee the existence of directional derivatives, while part (f) ensures that the second-order directional derivative is bounded from below by a quadratic function.[14]

Theorem 6.1. *Under Assumption* 6.2*, we have* $\widehat{\beta} - \beta = O_p(n^{-1/2})$ *for* $\widehat{\beta}$ *defined in Eq. (6.3).*

6.2 Empirical Illustration

Here we consider an empirical illustration, based on the aggregate panel multinomial choice model described above. We estimate a discrete choice demand model for bathroom tissue, using store/week-level scanner data from different branches of Dominicks supermarket.[15] The bathroom tissue category is convenient because there are relatively few brands of bathroom tissue, which simplifies the analysis. The data are collected at the store and week level, and report sales and prices of different brands of bathroom tissue. For each of 54 Dominicks stores, we aggregate the store-level sales of bathroom tissue up to the largest six brands, lumping the remaining brands into the seventh good (see Table 4).

[13] In the individual-level data setting, an analogue of Assumption 6.2(b) is implausible, as it would require a super-consistent nonparametric estimator of the conditional choice probability; for this reason, we conjecture that with individual-level data, the noise due to estimating the conditional choice probability would dominate and determine the rate. However, we have not found a way to handle this part of the noise.

[14] We use directional derivatives because our limiting criterion function is not fully differentiable to the second order. In particular, even though it is first-order differentiable, the first derivative has a kink.

[15] This dataset has previously been used in many papers in both economics and marketing; see a partial list at http://research.chicagobooth.edu/kilts/marketing-databases/dominicks/papers.

Table 4: Table of the 7 product-aggregates used in estimation.

	Products included in analysis
1	Charmin
2	White Cloud
3	Dominicks
4	Northern
5	Scott
6	Cottonelle
7	Other good (incl. Angelsoft, Kleenex, Coronet and smaller brands)

We form moment conditions based on cycles over weeks, for each store. In the estimation results below, we consider cycles of length 2. Since data are observed at the weekly level, we consider subsamples of 10 weeks or 15 weeks which were drawn at periodic intervals from the 1989-1993 sample period. After the specific weeks are drawn, all length-2 cycles that can be formed from those weeks are used.

Table 5: Summary Statistics

		min	max	mean	median	std.dev
10 week data	DEAL	0	1	.4350	0	.4749
	PRICE	.1776	.6200	.3637	.3541	.0876
	PRICE×DEAL	0	.6136	.1512	0	.1766
15 week data	DEAL	0	1	.4488	0	.4845
	PRICE	.1849	.6200	.3650	.3532	.0887
	PRICE×DEAL	0	.6091	.1644	0	.1888

We allow for store/brand level fixed effects and use the techniques developed in Section 3.1 to difference them out. Due to this, any time-invariant brand- or store-level variables will be subsumed into the fixed effect, leaving only explanatory covariates which vary both across stores and time. As such, we consider a simple specification with X^k = (PRICE, DEAL, PRICE*DEAL). PRICE is measured in dollars per roll of bathroom tissue, while DEAL is defined as whether a given brand was on sale in a given store-week.[16] Since any price discounts during a sale will be captured in the

[16]The variable DEAL takes the binary values $\{0, 1\}$ for products 1-6, but takes continuous values between 0 and 1 for product 7. The continuous values for product 7 stand for the average on-sale frequency of all the small brands included in the product-aggregate 7. This and the fact that PRICE is a continuous variable make the point

PRICE variable itself, DEAL captures any additional effects that a sale has on behavior, beyond price. Summary statistics for these variables are reported in Table 5.

Table 6: Point Estimates for Bathroom Tissue Choice Model

		10 week data	15 week data
β_1	deal	.1053	.0725
β_2	price	-.9720	-.9922
β_3	price×deal	-.2099	-.1017

The point estimates are reported in Table 6. One interesting observation from the table is that the sign of the interaction term is negative, indicating that consumers are more price sensitive when a product is on sale. This may be consistent with the story that the sale status draws consumers' attention to price (from other characteristics of the product).

7 Conclusions

In this paper we explored how the notion of cyclic monotonicity can be exploited for the identification and estimation of panel multinomial choice models with fixed effects. In these models, the social surplus (expected maximum utility) function is convex, implying that its gradient, which corresponds to the choice probabilities, satisfies cyclic monotonicity. This is just the appropriate generalization of the fact that the slope of a single-variate convex function is non-decreasing.

We establish sufficient conditions for point identification of the model parameters, and propose an estimator and show its consistency. Noteworthily, our moment inequalities are linear in the model parameters, so that the estimation procedure is a convex optimization problem, which has significant computational advantages. In ongoing work, we are considering the possible extension of these ideas to other models and economic settings.

identification condition, Assumption 3.3, hold.

References

[1] J. Abrevaya. Rank Estimation of a Generalized Fixed-effects Regression Model. *Journal of Econometrics*, 95: 1-23, 2000.

[2] D. Andrews and X. Shi. Inference based on conditional moment inequalities. *Econometrica*, 81: 609-666, 2013.

[3] H. Ahn, H. Ichimura, J. Powell, and P. Ruud. Simple Estimators for Invertible Index Models. Working paper, 2015.

[4] T. M. Apostol. *Calculus Volume 1: One-Variable Calculus, with an Introduction to Linear Algebra.* John Wiley and Sons, Inc. Second Edition. 1967.

[5] S. Berry and P. Haile. Identification in differentiated products markets using market-level data. *Econometrica*, 82: 1749-1797, 2014.

[6] S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, 65: 841-890, 1995.

[7] C. Bontemps, T. Magnac, and E. Maurin. Set Identified Linear Models. *Econometrica*, 80: 1129-1155, 2012.

[8] M. Browning. A Nonparametric Test of the Life-Cycle Rational Expectation Hypothesis. *International Economic Review*, 30:979-992, 1989.

[9] G. Chamberlain. Analysis of Variance with Qualitative Data. *Review of Economic Studies*, 47: 225-238, 1980.

[10] G. Chamberlain. Binary Response Models for Panel Data: Identification and Information. *Econometrica*, 78: 159-168, 2010.

[11] P. E. Cheng. Strong Consistency of Nearest Neighbor Regression Function Estimators. *Journal of Multivariate Analysis*, 15:63-72, 1984.

[12] V. Chernozhukov, S. Lee, and A. Rosen. Intersection Bounds: Estimation and Inference. *Econometrica*, 81: 667-737, 2013.

[13] J. Fox. Semi-parametric Estimation of Multinomial Discrete-Choice Models using a Subset of Choices. *RAND Journal of Economics*, 38: 1002-1029, 2007.

[14] J. Freyberger. Asymptotic Theory for Differentiated Product Demand Models with Many Markets. Working paper, 2013.

[15] J. Freyberger and J. Horowitz. Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation. Working paper, 2013.

[16] A. Han. Nonparametric Analysis of a Generalized Regression Model. *Journal of Econometrics*, 35:303-316, 1987.

[17] K. Hirano, G. W. Imbens, and G. Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71:1161-1189, 2003.

[18] S. Höderlein and H. White. Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects. *Journal of Econometrics*, 168:300-314, 2012.

[19] J. Horowitz. *Semi-parametric and Nonparametric Methods in Econometrics*. Springer-Verlag, 2009 (second edition).

[20] B. Honoré and E. Kyriazidou. Panel Discrete Choice Models with Lagged Dependent Variables. *Econometrica*, 68:839-874, 2000.

[21] B. Honoré and A. Lewbel. Semi-parametric binary choice panel data models without strictly exogenous regressors. *Econometrica*, 70:2053-2063, 2002.

[22] S. Khan, F. Ouyang, and E. Tamer. Adaptive Rank Inference in Semiparametric Multinomial Response Models. Working paper, 2016.

[23] S. Khan and E. Tamer Inference on Endogenously Censored Regression Models Using Conditional Moment Inequalities. *Journal of Econometrics*, 152:104-119, 2009.

[24] J. Kim and D. Pollard. Cube Root Asymptotics. *Annals of Statistics*, 18:191-219, 1990.

[25] R. H. Koning and G. Ridder. Discrete Choice and Stochastic Utility Maximization. *Econometric Journal*, 6: 1-27, 2003.

[26] A. Lewbel. Semi-parametric latent variable estimation with endogenous or mismeasured regressors. *Econometrica*, 66: 105-121, 1998.

[27] A. Lewbel. Semi-parametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97: 145-177, 2000.

[28] Q. Li and J. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

[29] C. F. Manski. The Maximum Score Estimation of the Stochastic Utility Model. *Journal of Econometrics*, 3:205–228, 1975.

[30] C. F. Manski. Semi-parametric Analysis of Random Effects Linear Models from Binary Panel Data. *Econometrica*, 55:357-362, 1987.

[31] D. L. McFadden. Modeling the Choice of Residential Location. In A. Karlgvist et. al., editors, *Spatial Interaction Theory and Residential Location*, North-Holland, 1978.

[32] D. L. McFadden. Economic Models of Probabilistic Choice. In C. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 1981.

[33] W. K. Newey and D. L. McFadden. Chapter 36 Large Sample Estimation and Hypothesis Testing. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics, Volume 4*, Elsevier, 1994.

[34] A. Pakes and J. Porter. Moment Inequalities for Semi-parametric Multinomial Choice with Fixed Effects. Working paper, Harvard University, 2013.

[35] J. Powell and P. Ruud. Simple Estimators for Semi-parametric Multinomial Choice Models. Working paper, 2008.

[36] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[37] R. Sherman. The Limiting Distribution of the Maximum Rank Correlation Estimator. *Econometrica*, 61: 123-137, 1993.

[38] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, 1996.

[39] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, Graduate Studies in Mathematics, Vol. 58, 2003.

A Appendix: Proofs

Proof of Lemma 2.1. (a) By the independence between $\mathbf{U}$ and $\boldsymbol{\epsilon}$, we have

$$\mathcal{W}(\mathbf{u}) = E\{\max_k [U^k + \epsilon^k] | \mathbf{U} = \mathbf{u}\} = E\{\max_k [u^k + \epsilon^k]\}. \tag{A.1}$$

This function is convex because $\max_k [u^k + \epsilon^k]$ is convex for all values of ϵ^k and the expectation operator is linear.

(b,c) Without loss of generality, we focus on the differentiability with respect to u^K. Let $(u_*^1, \ldots, u_*^K)$ denote an arbitrary fixed value of $(U^1, \ldots, U^K)$, and let $u_*^0 = 0$. It suffices to show that $\lim_{\eta \to 0} [\mathcal{W}(u_*^1, \ldots, u_*^K + \eta) - \mathcal{W}(u_*^1, \ldots, u_*^K)]/\eta$ exists. We show this using the bounded convergence theorem. First observe that

$$\frac{\mathcal{W}(u_*^1, \ldots, u_*^K + \eta) - \mathcal{W}(u_*^1, \ldots, u_*^K)}{\eta} = E\left[\frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{\epsilon})}{\eta}\right], \tag{A.2}$$

where $\Delta(\eta, \mathbf{u}_*, \boldsymbol{\epsilon}) = \max\{u_*^1 + \epsilon^1, \ldots, u_*^K + \eta + \epsilon^K\} - \max\{u_*^1 + \epsilon^1, \ldots, u_*^K + \epsilon^K\}$. Consider an arbitrary value $\boldsymbol{e}$ of $\boldsymbol{\epsilon}$ and $e^0 = 0$. If $e^K + u_*^K > \max_{k=0,\ldots,K-1}[u_*^k + e^k]$, for η close enough to zero, we have

$$\frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{e})}{\eta} = \frac{(u_*^K + \eta + e^K) - (u_*^K + e^K)}{\eta} = 1. \tag{A.3}$$

Thus,

$$\lim_{\eta \to 0} \frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{e})}{\eta} = 1. \tag{A.4}$$

On the other hand, if $e^K + u_*^K < \max_{k=0,\ldots,K-1}[u_*^k + e^k]$, then for η close enough to zero, we have

$$\frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{e})}{\eta} = \frac{0}{\eta} = 0. \tag{A.5}$$

Thus,

$$\lim_{\eta \to 0} \frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{e})}{\eta} = 0. \tag{A.6}$$

Because $\boldsymbol{\epsilon}$ has a continuous distribution, we have $\Pr(\epsilon^K + u_*^K = \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]) = 0$. Therefore, almost surely,

$$\lim_{\eta \to 0} \frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{\epsilon})}{\eta} = 1\{\epsilon^K + u_*^K > \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]\}. \tag{A.7}$$

Also, observe that

$$\left|\frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{\epsilon})}{\eta}\right| \leq \left|\frac{u_*^K + \eta + \epsilon^K - (u_*^K + \epsilon^K)}{\eta}\right| = 1 < \infty. \tag{A.8}$$

Thus, the bounded convergence theorem applies and yields

$$\lim_{\eta \to 0} E\left[\frac{\Delta(\eta, \mathbf{u}_*, \boldsymbol{\epsilon})}{\eta}\right] = E[1\{\epsilon^K + u_*^K > \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]\}] = p^K(\mathbf{u}). \tag{A.9}$$

This shows both part (b) and part (c).

Part (d) is a direct consequence of part (c) and Proposition 1. □

Proof of Theorem 3.1. To prove Theorem 3.1, we first prove the following lemma.

Define the convex conic hull of $\mathcal{G}$ as:

$$\text{coni}(\mathcal{G}) = \left\{ \sum_{\ell=1}^{L} \lambda_\ell g_\ell \middle| \, g_\ell \in \mathcal{G}, \lambda_\ell \in R, \lambda_\ell \geq 0; \ell, L = 1, 2, \dots \right\}. \tag{A.10}$$

Lemma A.1. *Suppose that the set* $\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \text{coni}(\mathcal{G})$*, then* $Q(\beta) = 0$*, and* $Q(b) > 0$ *for all* $b \in \{b' \in R^{d_x} : \|b'\| = 1\}$ *such that* $b \neq \beta$*.*

Proof of Lemma A.1. The result $Q(\beta) = 0$ is straightforward due to Equation (3.8). We next show that for any $b \neq \beta$ and $\|b\| = 1$, $Q(b) > 0$.

Suppose not, that is, suppose that $Q(b) = 0$. Then we must have $b'g \geq 0$ for all $g \in \mathcal{G}$ because if not, there must be a subset $\mathcal{G}_0$ of $\mathcal{G}$ such that $\Pr(\mathbf{g} \in \mathcal{G}_0) > 0$ and $b'g < 0 \ \forall g \in \mathcal{G}_0$ which will imply $Q(b) > 0$. Now that $b'g \geq 0$ for all $g \in \mathcal{G}$, it must be that

$$b'g \geq 0 \ \forall g \in \text{coni}(\mathcal{G}). \tag{A.11}$$

This implies that

$$\text{coni}(\mathcal{G}) \subseteq \{g \in R^{d_x} : b'g \geq 0\}. \tag{A.12}$$

Combining that with the condition of the lemma, we have

$$\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \{g \in R^{d_x} : b'g \geq 0\}. \tag{A.13}$$

This implies that $\beta = b$, which contradicts the assumption that $b \neq \beta$. This concludes the proof of the lemma. □

Now we prove Theorem 3.1 using the lemma we just proved. By the lemma, it suffices to show that

$$\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \text{coni}(\mathcal{G}). \tag{A.14}$$

We break the proof into two cases depending on whether assumption (3.3) or (3.4) is assumed to hold.

Under assumption 3.3 (continuous covariates). Suppose that Assumption 3.3 holds. Below we establish two facts:
(i) $\{g \in R^{d_x} : \beta' g > 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in G, \beta' g > 0\}$; and
(ii) $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in G, \beta' g > 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$.
These two facts (i) and (ii) together immediately imply that

$$\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\} \subseteq \text{coni}(\mathcal{G}) \tag{A.15}$$

where the last subset inclusion follows from the definition of coni$(\cdot)$. This proves (A.14).

To show (i), consider an arbitrary point $g_0 \in R^{d_x}$ such that $\beta' g_0 > 0$. Then by Assumption 3.3, there exists a $\lambda \geq 0$ and a $g \in G$ such that $\lambda g = g_0$. Because $\beta' g_0 > 0$, we must have $\lambda \beta' g > 0$, and thus $\beta' g > 0$. That implies, $g \in \tilde{G}$. This shows result (i).

To show (ii), consider an arbitrary point in $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in G, \beta' g > 0\}$. Then this point can be written as $\lambda^* g^*$ where λ^* is a scalar such that $\lambda^* \geq 0$ and g^* is an element in G such

that $\beta' g^* > 0$. By the definition of G, we have either $g^* \in \text{supp}(\pm \Delta X_i^k | \Delta X_i^{-k} = 0)$, for some $k \in \{1, \ldots, K\}$, or $g^* \in \text{supp}(\pm \Delta X_i^1 | \Delta X_i^k = \Delta X_i^1 \ \forall k)$. We discuss these two cases separately.

First, suppose without loss of generality $g^* \in \text{supp}(\Delta X_i^k | \Delta X_i^{-k} = 0)$ for some $k \in \{1, \ldots, K\}$. Then there exists x_*^k and $x_\dagger^k$ such that $x_*^k - x_\dagger^k = g^*$ and $(x_*^k, x_\dagger^k)$ is in the conditional support of (X_{i2}^k, X_{i1}^k) given $\Delta \mathbf{X}_i^{-k} = \mathbf{0}$. By the definition of $\mathcal{G}$ and Assumption 3.2(b), we have

$$\{E[\Delta Y_i^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = \mathbf{0}, X_{i1}^k = x_\dagger^k] g^*\} \in \mathcal{G}. \tag{A.16}$$

Note that Assumption 3.2(b) is used to guarantee that $E[\Delta Y_i^k | X_{i2}^k = x_*^k, \Delta X_i^{-k} = 0, X_{i1}^k = x_\dagger^k](x_*^k - x_\dagger^k)$ is a continuous function and thus maps the support of (X_{i2}^k, X_{i1}^k) into the support of $E[\Delta Y_i^k | X_{i2}^k, \Delta \mathbf{X}_{it}^{-k} = \mathbf{0}, X_{i1}^k] \Delta X_i^k$. Below we show that

$$a := E[\Delta Y_i^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = 0, X_{i1}^k = x_\dagger^k] > 0. \tag{A.17}$$

This and (A.16) together imply that

$$\lambda^* g^* = (\lambda^* a^{-1}) a g^* \in \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\} \tag{A.18}$$

This shows result (ii).

The result in (A.17) follows from the derivation:

$$\begin{aligned}
&E[Y_{i2}^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = 0, X_{i1}^k = x_\dagger^k] \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i2}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i2}^{k'} + A_i^{k'} + \epsilon_{i2}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i2}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i1}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&> \Pr\left(\beta' x_\dagger^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i1}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= E[Y_{i1}^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = 0, X_{i1}^k = x_\dagger^k],
\end{aligned} \tag{A.19}$$

where the first and the last equalities hold by the specification of the multinomial choice model, the second equality holds by Assumption 3.1(a), the third equality is obvious from the conditioning event, and the inequality holds by Assumption 3.2(a) and $\beta'(x_*^k - x_\dagger^k) > 0$.

Second, suppose instead, and without loss of generality, $g^* \in \text{supp}(\Delta X_i^1 | \Delta X_i^k = \Delta X_i^1 \ \forall k)$. Then there exists $(x_*^k, x_\dagger^k)_{k=1}^K$ in the support of (X_{i2}^k, X_{i1}^k) such that $g^* = x_*^k - x_\dagger^k$ for all $k = 1, \ldots, K$. By the definition of $\mathcal{G}$ and Assumption 3.2(b), the following vector belongs to $\mathcal{G}$:

$$-E[\Delta Y_i^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K] g. \tag{A.20}$$

where $\Delta Y_i^0 = Y_{i2}^0 - Y_{i1}^0$ and $Y_{it}^0 = 1 - \sum_{k=1}^K Y_{it}^k$. Below we show that

$$a := -E[\Delta Y_{it}^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K] > 0. \tag{A.21}$$

The rest of the prove of result (ii) is the same as that in the first case above.

Inequality (A.21) follows from the derivation:

$$
\begin{aligned}
&E[Y_{i2}^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K] \\
&= \Pr\left(\max_{k=1,\ldots,K} \beta' x_*^k + A_i^k + \epsilon_{i2}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&= \Pr\left(\max_{k=1,\ldots,K} \beta' x_*^k + A_i^k + \epsilon_{i1}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&< \Pr\left(\max_{k=1,\ldots,K} \beta' x_\dagger^k + A_i^k + \epsilon_{i1}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&= E[Y_{i1}^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K],
\end{aligned} \tag{A.22}
$$

where the arguments for each steps are the same as those for the corresponding steps in (A.19).

Under assumption 3.4: discrete covariates. Suppose that Assumption 3.4 holds. It has been shown in the continuous covariate case above that $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in G, \beta' g > 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$ under Assumptions 3.1(a) and 3.2. That implies

$$
\text{coni}(\{g \in G : \beta' g \geq 0\}) \subseteq \text{coni}(\mathcal{G}). \tag{A.23}
$$

Below we show that

$$
\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \text{coni}(\{g \in G : \beta' g \geq 0\}). \tag{A.24}
$$

This combined with (A.23) proves (A.14) and thus proves the theorem.

Now we show (A.24). Suppose without loss of generality that $\beta_{j^*} > 0$. Let $\tilde{G}^0 = \{g \in R^{d_x} : g_{-j} \in G_{-j}^0, g_{j^*} > -\beta_{-j^*}' g_{-j^*} / \beta_{j^*}\}$, where $\beta_{-j^*} = (\beta_1, \ldots, \beta_{j^*-1}, \beta_{j^*+1}, \ldots, \beta_{d_x})'$. By Assumption 3.4(a), we have that

$$
\tilde{G}^0 \subseteq \{g \in G : \beta' g \geq 0\}. \tag{A.25}
$$

Consider an arbitrary point $g_0 \in \{g \in R^{d_x} : \beta' g \geq 0\}$. Then, $g_{0,j^*} > -g_{0,-j^*}' \beta_{-j^*} / \beta_{j^*}$. That means

$$
d := g_{0,j^*} + g_{0,-j^*}' \beta_{-j^*} / \beta_{j^*} > 0. \tag{A.26}
$$

By Assumption 3.4(b), $G_{-j^*}^0$ spans $R^{d_x - 1}$, and is symmetric about the origin. Thus, $G_{-j^*}^0$ spans $R^{d_x - 1}$ with nonnegative weights. Then, there exists a positive integer M, weights $c_1, \ldots, c_M > 0$, and $g_{1,-j^*}, \ldots, g_{M,-j^*} \in G_{-j^*}^0$ such that $g_{0,-j^*} = \sum_{m=1}^M c_m g_{m,-j^*}$.

Let $g_{m,j^*} = \left(d / \sum_{m=1}^M c_m \right) - \left(g_{m,-j^*}' \beta_{-j^*} / \beta_{j^*} \right)$ for $m = 1, \ldots, M$. Let g_m be the vector whose j^*th element is g_{m,j^*} and who with the j^* element removed is $g_{m,-j^*}$, for $m = 1, \ldots, M$. Then $g_m \in \tilde{G}^0$ for $m = 1, \ldots, M$ because $g_{m,-j^*} \in G_{-j}^0$ by construction and $g_{m,j^*} > -g_{m,-j^*}' \beta_{-j^*} / \beta_{j^*}$ due to $d > 0$. Also it is easy to verify that $g_0 = \sum_{m=1}^M c_m g_m$. Thus, $g_0 \in \text{coni}(\tilde{G}^0)$. Subsequently, by (A.25)

$$
g_0 \in \text{coni}(\{g \in G : \beta' g \geq 0\}). \tag{A.27}
$$

Therefore, (A.24) holds. □

Proof of Theorem 4.1. For any $b \in R^{d_x}$, let $\|b\|_\infty = \max_{j=1,\ldots,J} |b_j|$. Below we show that

$$\overline{\beta} \to_p \beta/\|\beta\|_\infty. \tag{A.28}$$

This implies that $\widehat{\beta} \to_p \beta$ because $\widehat{\beta} = \overline{\beta}/\|\overline{\beta}\|$ and the mapping $f : \{b \in R^{d_x} : \|b\|_\infty = 1\} \to \{b \in R^{d_x} : \|b\| = 1\}$ such that $f(b) = b/\|b\|$ is continuous.

Now we show Eqn. (A.28). Let

$$Q(b) = E\left[b'(\Delta \mathbf{X}_i)\left(\Delta \mathbf{p}(\mathbf{X}_{i1}, \mathbf{X}_{i2})\right)\right]_-. \tag{A.29}$$

Under Assumption 3.1, the identifying inequalities (3.7) hold, which implies that

$$Q(\beta) = Q(\beta/\|\beta\|_\infty)) = 0. \tag{A.30}$$

Consider any b such that $\|b\|_\infty = 1$ and $b \neq \beta/\|\beta\|_\infty$. We have $b/\|b\| \neq \beta/\|\beta\|$ because the function $f(b) = b/\|b\| : \{b \in R^{d_x} : \|b\|_\infty = 1\} \to \{b \in R^{d_x} : \|b\| = 1\}$ is one-to-one. Thus, for such a b, Theorem 3.1 implies that,

$$Q(b) > 0. \tag{A.31}$$

This, the continuity of $Q(b)$, and the compactness of the parameter space $\{b \in R^{d_x} : \|b\|_\infty = 1\}$ together imply that, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that,

$$\inf_{b \in R^{d_x} : \|b\|_\infty = 1, \|b-\beta\| > \varepsilon} Q(b) \geq \delta. \tag{A.32}$$

Next we show the uniform convergence of $Q_n(b)$ to $Q(b)$. Combining (A.31) and the uniform convergence, one can show the consistency of $\widehat{\beta}$ using standard consistency arguments in, e.g., Newey and McFadden (1994)).

Now we show the uniform convergence of $Q_n(b)$ to $Q(b)$. That is, we show that

$$\sup_{b \in R^{d_x} : \|b\|_\infty = 1} |Q(b) - Q_n(b)| \to_p 0. \tag{A.33}$$

First, we show the stochastic equicontinuity of $Q_n(b)$. For any $b, b^* \in R^{d_x}$ such that $\|b\|_\infty = \|b^*\|_\infty = 1$, consider the following derivation:

$$\begin{aligned}
|Q_n(b) - Q_n(b^*)| &\leq n^{-1} \sum_{i=1}^n \left|(b - b^*)'(\Delta \mathbf{X}_i)\left(\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2})\right)\right| \\
&\leq n^{-1} \sum_{i=1}^n \|b - b^*\| \|(\Delta \mathbf{X}_i)\left(\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2})\right)\| \\
&\leq 2n^{-1} \sum_{i=1}^n \|\Delta \mathbf{X}_i\| \|b - b^*\|.
\end{aligned} \tag{A.34}$$

Therefore, for any fixed $\varepsilon > 0$, we have

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \Pr\left(\sup_{b, b^* \in R^{d_x}, \|b\|_\infty = \|b^*\|_\infty = 1, \|b - b^*\| \leq \delta} |Q_n(b) - Q_n(b^*)| > \varepsilon \right)$$

$$
\begin{aligned}
&\leq \lim_{\delta \downarrow 0} \limsup_{n \to \infty} \Pr\left(2\delta n^{-1} \sum_{i=1}^{n} \|\Delta \mathbf{X}_i\| > \varepsilon \right) \\
&\leq \lim_{\delta \downarrow 0} \limsup_{n \to \infty} \Pr\left(2 n^{-1} \sum_{i=1}^{n} \|\Delta \mathbf{X}_i\| > \varepsilon/\delta \right) \\
&= 0, \qquad \text{(A.35)}
\end{aligned}
$$

where the first inequality holds by (A.34) and the equality holds by Assumption 4.1(b). This shows the stochastic equicontinuity of $Q_n(b)$.

Given the stochastic equicontinuity $Q_n(b)$ and the compactness of $\{b \in R^{d_x} : \|b\|_\infty = 1\}$, to show (A.33), it suffices to show that for all $b \in R^{d_x} : \|b\|_\infty = 1$, we have

$$
Q_n(b) \to_p Q(b). \qquad \text{(A.36)}
$$

For this purpose, let

$$
\widetilde{Q}_n(b) = n^{-1} \sum_{i=1}^{n} \left[(b' \Delta \mathbf{X}_i)(\Delta \mathbf{p}(\mathbf{X}_{i1}, \mathbf{X}_{i2})) \right]_- . \qquad \text{(A.37)}
$$

By Assumption 4.1(b) and the law of large numbers, we have $\widetilde{Q}_n(b) \to_p Q(b)$. Now we only need to show that $|\widetilde{Q}_n(b) - Q_n(b)| \to_p 0$. But that follows from the derivation:

$$
\begin{aligned}
&|\widetilde{Q}_n(b) - Q_n(b)| \\
&\leq n^{-1} \sum_{i=1}^{n} \left| (b' \Delta \mathbf{X}_{i1})(\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \Delta \mathbf{p}(\mathbf{X}_{i1}, \mathbf{X}_{i2})) \right| , \\
&\leq 2 \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in \text{supp}(\mathbf{X}_{it})} \sup_{t=1,2} \|\hat{\mathbf{p}}_t(\mathbf{x}_1, \mathbf{x}_2) - \mathbf{p}_t(\mathbf{x}_1, \mathbf{x}_2)\| \right) \left(n^{-1} \sum_{i=1}^{n} \|b' \mathbf{X}_{i1} - b' \mathbf{X}_{i2}\| \right) , \\
&\to_p 0, \qquad \text{(A.38)}
\end{aligned}
$$

where the convergence holds by Assumptions 4.1(a)-(b). Therefore the theorem is proved. □

Proof of Theorem 6.1. First we define the limiting version of $Q_n(b)$ as $n \to \infty$:

$$
Q(b) = E\left\{ [(b' \Delta \mathbf{X}_c) E(\Delta \mathbf{Y}_{ic} | \eta_c)]_- \right\} . \qquad \text{(A.39)}
$$

Let B_δ stand for the set $\{b \in R^{d_x} : b_1 = 1, \|b - \beta\| \leq \delta\}$. Below we show the folowing results: given a positive number δ, for all $b \in B_\delta$,

(i) for all $\eta > 0$ (regardless of how small), $|Q_n(b) - Q(b) - (Q_n(\beta) - Q(\beta))| \leq \eta \|b - \beta\|^2 + O_p(n^{-1})$, and

(ii) $Q(b) - Q(\beta) \geq c_2 \|b - \beta\|^2/2$, where c_2 is the constant in Assumption 6.2(f).

Choose $\eta < c_2/2$. Results (i) and (ii) together with Assumption 6.2(c) imply that, with probability approaching one,

$$
c_2 \|\widehat{\beta} - \beta\|^2/2 \leq \eta \|\widehat{\beta} - \beta\|^2 + O_p(n^{-1}) + Q_n(\widehat{\beta}) - Q_n(\beta)
$$

$$\leq \eta\|\widehat{\beta} - \beta\|^2 + O_p(n^{-1}). \tag{A.40}$$

where the second inequality holds because $\widehat{\beta}$ minimizes $Q_n(\cdot)$ and hence $Q_n(\widehat{\beta}) - Q_n(\beta) \leq 0$. Thus,

$$(c_2/2 - \eta)\|\widehat{\beta} - \beta\|^2 \leq O_p(n^{-1}). \tag{A.41}$$

This proves the theorem because $(c_2/2 - \eta) > 0$ by design.

Now we show result (i). Note that

$$\begin{aligned} Q_n(b) - Q(b) = n^{-1}\sum_{c=1}^{n}[b'(\Delta\mathbf{X}_c)(\Delta\mathbf{S}_c)]_- - E\left\{[b'(\Delta\mathbf{X}_c)(\Delta\mathbf{S}_c)]_-\right\} \\ + E\left\{[b'(\Delta\mathbf{X}_c)(\Delta\mathbf{S}_c)]_-\right\} - E\left\{[b'(\Delta\mathbf{X}_c)E(\Delta\mathbf{Y}_{ic}|\eta_c)]_-\right\}. \end{aligned} \tag{A.42}$$

The first two summands on the right hand-side together form an empirical process which we now call $\nu_n(b)$ and will analyze later. The absolute value of the rest of the right hand-side is bounded by

$$\begin{aligned} &E\left\{\|b'(\Delta\mathbf{X}_c)\|\|(\Delta\mathbf{S}_c) - E(\Delta\mathbf{Y}_{ic}|\eta_c)\|\right\} \\ &\leq 2\max_{t=1,2}(E\|\mathbf{S}_{ct} - E(\mathbf{Y}_{ict}|\eta_c)\|^2)^{1/2}(E\|b'(\Delta\mathbf{X}_c)\|^2)^{1/2} \\ &\leq O(n^{-1})\|b\|(E\|\mathrm{vec}(X_{c2} - X_{c1})\|^2)^{1/2}, \end{aligned} \tag{A.43}$$

where the first inequality holds by the Cauchy-Schwarz inequality, the second inequality holds by Assumption 6.2(b) and the Cauchy-Schwarz inequality. The last line is $o_p(n^{-1})$ uniformly over a $o_p(1)$ neighborhood of β by Assumption 6.2(a). Therefore, we have, uniformly over a $o_p(1)$ neighborhood of β,

$$Q_n(b) - Q(b) - (Q_n(\beta) - Q(\beta)) = \nu_n(b) - \nu_n(\beta) + O(n^{-1}). \tag{A.44}$$

We now bound $\nu_n(b) - \nu_n(\beta)$. Let V_c denote $(\Delta\mathbf{X}_c)(\Delta\mathbf{S}_c)$, and let the space of V_c be denoted by $\mathcal{V}$.

We first show that the class of functions $\mathcal{F} = \{f : \mathcal{V} \to R | f(v) = [b'v]_-, b \in R^{d_x}, b_1 = 1\}$ is a Vapnik-Cervonenkis (VC)-subgraph class of functions. To begin, observe that a similar but different class of functions with $\mathcal{F}$: $\mathcal{F}_0 = \{f : \mathcal{V} \to R | f(v) = -b'v\}$ is a VC-subgraph of VC-index at most $d_x + 1$ because the space is a vector space of dimension equal to the dimension of the set $\{b \in R^{d_x} : b_1 = 1\}$. That implies that the collection of subgraphs $\{(v, a) : -b'v > a\}$ of functions in $\mathcal{F}_0$ forms a VC-class of sets of dimension at most $d_x + 1$. We use this to show that the collection of subgraphs of functions in $\mathcal{F}$ is also a VC-class with finite dimension. The subgraph of a function in $\mathcal{F}$ is

$$S(b) = \{(v, a) : a < [b'v]_-\} = \{(v, a) : a < 0\} \cup \{(v, a) : a \geq 0, -b'v > a\}. \tag{A.45}$$

Consider m points $(v_1, a_1), \ldots, (v_m, a_m)$. In order for $\{S(b) : b \in R^{d_x}, b_1 = 1\}$ to shatter these m points, there can at most be one $j \in \{1, \ldots, m\}$ such that $a_j < 0$ because $S(b)$ with different b cannot pick out two different points with $a < 0$. Suppose without loss of generality that $a_1 < 0$. Then, the collection of sets $\{\{(v, a) : -b'v > a\} : b \in R^{d_x}, b_1 = 1\}$ must shatter the set $\{(v_2, a_2), \ldots, (v_m, a_m)\}$. But this collection of sets is the collection of subgraphs of functions in $\mathcal{F}_0$, which is of VC-dimension

at most $d_x + 1$. Therefore, $m - 1$ can at most be $d_x + 1$. This implies that $m \leq d_x + 2$. Thus, $\mathcal{F}$ is a VC-subgraph with VC-index at most $d_x + 2$.

Next define

$$\mathcal{F}_\delta = \{f : \mathcal{V} \to R | f(v) = [v'b]_- - [v'\beta]_-, b \in R^{d_x}, b_1 = 1, \|b - \beta\| \leq \delta\}. \tag{A.46}$$

This collection of functions is a VC-class with the same VC-index as $\mathcal{F}$ due to Lemma 2.6.18 of van der Vaart and Wellner (1996). Consider the envelope function $F_\delta(v) = \|v\|\delta$. Then, Theorem 2.6.7 of van der Vaart of Wellner (1996) gives the polynomial bound on the covering number of $\mathcal{F}_\delta$:

$$N(\varepsilon E_Q \|V_c\|^2 \delta^2, \mathcal{F}_\delta, L_2(Q)) \leq C\varepsilon^{-2d_x - 2}, \tag{A.47}$$

where $N(\varepsilon E_Q \|V_c\|^2 \delta^2, \mathcal{F}_\delta, L_2(Q))$ is the covering number of $\mathcal{F}_\delta$ by $L_2(Q)$ balls of radius $\varepsilon E_Q \|V_c\|^2 \delta^2$, and Q is an arbitrary probability measure on $\mathcal{V}$, and C is a universal constant that depends only on d_x. Next we apply Theorem 2.14.1 of van der Vaart and Wellner (1996) to bound $\nu_n(b) - \nu_n(\beta)$:

$$\begin{aligned}
&E\left\{\sup_{b \in R^{d_x} : b_1 = 1, \|b-\beta\| \leq \delta} |\nu_n(b) - \nu_n(\beta)|^2\right\} \\
&\leq C \sup_Q \int_0^1 \sqrt{1 + \log N(\varepsilon E_Q \|V_c\|^2 \delta^2, \mathcal{F}_\delta, L_2(Q))} d\varepsilon \times E\|V_c\|^2 \delta^2 / n \\
&= C \int_0^1 \sqrt{1 + \log C - (2d_x + 2) \log \varepsilon} d\varepsilon \times E\|V_c\|^2 \delta^2 / n,
\end{aligned} \tag{A.48}$$

where the $C's$ are universal constants which may not be the same each time it appears. A change of variable technique can be used to show that the integral is finite. That combined with $E(\|\text{vec}(\mathbf{X}_{ct})\|^2) < \infty$ (Assumption 6.2(a)) implies that

$$E\left\{\sup_{b \in R^{d_x} : b_1 = 1, \|b-\beta\| \leq \delta} |\nu_n(b) - \nu_n(\beta)|^2\right\} \leq C\delta^2 / n. \tag{A.49}$$

Using this and the arguments used in the proof of Lemma 4.1 of Kim and Pollard (1990), we can show that for arbitrarily small η, we have for all b such that $b_1 = 1, \|b - \beta\| \leq \delta$,

$$|\nu_n(b) - \nu_n(\beta)| \leq \eta \|b - \beta\|^2 + O_p(n^{-1}). \tag{A.50}$$

This combined with (A.44) shows result (i).

Finally, we show result (ii). Consider any $h = (0, \tilde{h}')' \in R^{d_x}$ such that $\|h\| = 1$. Consider $Q(\beta + hz)$ as a function of z at a given β and h value. Below we show that for all $z \in [0, c_1)$,

$$\frac{\partial Q(\beta + zh)}{\partial z} = -E[\mathbf{W}_c' h 1\{\mathbf{W}_c'(\beta + zh) < 0\}], \tag{A.51}$$

and that this first derivative is continuous in z. Below we also show that for all $z \in (0, c_1)$,

$$\frac{\partial^2 Q(\beta + zh)}{\partial z^2} = E[(\tilde{\mathbf{W}}_c' \tilde{h})^2 f_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c' \tilde{\beta} - z\tilde{\mathbf{W}}_c' \tilde{h} | \tilde{\mathbf{W}}_c) 1(\tilde{\mathbf{W}}_c' \tilde{h} < 0)]. \tag{A.52}$$

Given those, for any $z \in (0, c_1)$, a Taylor expansion with Lagrangian remainder applies and gives[17]

$$\begin{aligned} Q(\beta + hz) &= Q(\beta) + \frac{\partial Q(\beta + 0h)}{\partial t} z + 2^{-1} \frac{\partial^2 Q(\beta + \tau h)}{\partial z^2} z^2 \\ &= 0 + 2^{-1} z^2 \tilde{h}' E[\tilde{\mathbf{W}}_c \tilde{\mathbf{W}}_c' f_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c' \tilde{\beta} - \tau \tilde{\mathbf{W}}_c' \tilde{h} | \tilde{\mathbf{W}}_c) 1(\tilde{\mathbf{W}}_c' \tilde{h} < 0)] \tilde{h} \\ &\geq c_2 z^2 \|\tilde{h}\| = c_2 z^2 \end{aligned} \tag{A.53}$$

for a τ in between 0 and z, where the inequality holds by Assumption 6.2(f). For an arbitrary $b \in B_{c_1}$, let $z = \|b - \beta\|$, and let $h = (b - \beta)/\|b - \beta\|$. The above display shows Result (ii).

Now we derive the first derivative of $Q(\beta + hz)$ with respect to z. Its first derivative equals the limit of the following quantity as $\tau \to z$, if the limit exists:

$$\frac{E([\mathbf{W}_c'(\beta + \tau h)]_-) - E([\mathbf{W}_c'(\beta + zh)]_-)}{\tau - z} \tag{A.54}$$

By Assumption 6.2(d), we have, for small enough z ($z < c_1$) with probability one

$$\lim_{\tau \to z} \frac{[\mathbf{W}_c'(\beta + zh + (\tau - z)h)]_- - [\mathbf{W}_c'(\beta + zh)]_-}{\tau - z} = -\mathbf{W}_c' h 1\{\mathbf{W}_c'(\beta + zh) < 0\}. \tag{A.55}$$

Also observe that

$$\begin{aligned} \left| \frac{[\mathbf{W}_c'(\beta + \tau h)]_- - [\mathbf{W}_c'(\beta + zh)]_-}{\tau - z} \right| &\leq \left| \frac{\mathbf{W}_c'(\beta + \tau h) - [\mathbf{W}_c'(\beta + zh)}{\tau - z} \right| \\ &= |\mathbf{W}_c' h|. \end{aligned} \tag{A.56}$$

Assumption 6.2(a) implies that the right hand-side has finite first moment. Equations (A.55) and (A.56) together combined with the dominated convergence theorem imply that,

$$\lim_{\tau \to z} E\left(\frac{[\mathbf{W}_c'(\beta + \tau h)]_- - [\mathbf{W}_c'(\beta + zh)]_-}{\tau - z} \right) = -E[\mathbf{W}_c' h 1\{\mathbf{W}_c'(\beta + zh) < 0\}]. \tag{A.57}$$

This shows (A.51). The derivative is continuous in z by a similar application of the dominated convergence theorem.

Next we derive the second derivative at $z \in (0, c_1)$. It is convenient to write $\partial Q(\beta + zh)/\partial z$ as

$$\frac{\partial Q(\beta + zh)}{\partial z} = -E\left[\tilde{\mathbf{W}}_c' \tilde{h} F_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c'(\tilde{\beta} + z\tilde{h}) | \tilde{\mathbf{W}}_c) \right] \tag{A.58}$$

The derivative of of this equals the limit of the following quantity as $\tau \to z$, if the limit exists:

$$E\left(\frac{\tilde{\mathbf{W}}_c' \tilde{h} F_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c'(\tilde{\beta} + \tau\tilde{h}) | \tilde{\mathbf{W}}_c) - \tilde{\mathbf{W}}_c' \tilde{h} F_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c'(\tilde{\beta} + z\tilde{h}) | \tilde{\mathbf{W}}_c)}{\tau - z} \right). \tag{A.59}$$

Under Assumption 6.2(e), the limit of the quantity inside the large brackets exists almost surely and equals

$$-(\tilde{\mathbf{W}}_c' \tilde{h})^2 f_{\mathbf{W}_c^1 | \tilde{\mathbf{W}}_c}(-\tilde{\mathbf{W}}_c' \tilde{\beta} - z\tilde{\mathbf{W}}_c' \tilde{h} | \tilde{\mathbf{W}}_c) 1(z\tilde{\mathbf{W}}_c' \tilde{h} < 0). \tag{A.60}$$

Also, under Assumption 6.2(e), the absolute value of the quantity inside the large brackets in (A.59) is bounded above by $|C(\tilde{\mathbf{W}}_c' \tilde{h})^2|$, which has finite first moment by Assumption 6.2(a) and the fact that $\|\tilde{h}\| = 1$. Therefore, the bounded convergence theorem applies and shows that the limit of (A.59) exists and equals the expectation of (A.60). This concludes the proof of (A.52). □

[17] See, for example, Apostol (1967, Section 7.7).

B Appendix: Primitive necessary condition for point identification

In this section we characterize a primitive necessary condition for point identification, in the special case of a binary choice model.[18]

In the binary choice case, it is without loss to consider only cycles of length 2. Moreover, because $K = 1$, there is no need for the bold font on X_{it}, ϵ_{it}, A_i, v, and a. Similarly, there is also no need for the choice index superscript on these symbols. Thus, we omit them in this section.

Consider the G set defined in Section 3.2 and specialized to the binary case. Theorem B.1 below is the main result of this section. It shows that, if one regressor has finite support and all other regressors have bounded support, then point identification cannot be achieved at all values of β.

Assumption B.1. *For some $j = 1, \ldots, d_x$,* (a) *G_j is a finite set, and*
(b) *G_{-j} is a bounded set.*

Theorem B.1 (**Necessary conditions for point identification).** *Under Assumptions* 3.1(a)-(b) *and* 3.2*, if Assumption* B.1 *holds, then it is not always true that $Q(b) > 0$ for all $b \in R^{d_x}$ such that $\|b\| = 1$ and $b \neq \beta$.*

Remark. According to the Theorem B.1, if one coordinate of $X_{is} - X_{it}$ has finite support for all s, t, then another coordinate of it must have unbounded support for some pair (s, t). The variable $X_{j,is} - X_{j,it}$ may have finite support, either when $X_{j,it}$ has finite support, or when the change of $X_{j,it}$ across time periods is restricted to a few grids. When that is the case, point identification requires that another regressor, say, $X_{j',it}$ to change unboundedly as t changes.

Theorem B.1 does not imply that β can never be point identified (up to scale normalization). There can be β values such that, when the population is generated from the model specified in (1.1) and (1.2) with β being that value, $Q(b) > 0$ for all $b \in \{b \in R^{d_x} : \|b\| = 1\}$ such that $b \neq \beta$. In other words, under the conditions of the theorem, point identification may be achieved in part of the parameter space, but not on the whole space of β. ■

Proof of Theorem B.1. It suffices to find at least one β value that generates a population for which point identification fails. Below we find such a value among β's that satisfy $\beta_j > 0$, $\beta_{j^*} > 0$ for some $j^* \neq j$, and $\beta_{j'} = 0$ for $j' \neq j, j^*$. It is useful to note that G is symmetric about the origin by definition. So are $G_{j'}$'s for all $j' = 1, \ldots, d_x$.

We discuss two cases. In the first case, $G_j \cap (-\infty, 0) = \emptyset$. Then $G_j = \{0\}$ because it is symmetric about the origin. Then $\mathcal{G}$ is contained in the subspace $\{g \in R^{d_x} : g_j = 0\}$. Let b^* be equal to β except that $b^*_j = 0$, and let $b = b^*/\|b^*\|$. Then $(b^*)'g = \beta' g$ for all $g \in \mathcal{G}$. This implies that $b'g \geq 0$ for all $g \in \mathcal{G}$, and thus $Q(b) = Q(\beta) = 0$.

In the second case, $G_j \cap (-\infty, 0) \neq \emptyset$. Assumption B.1(a) implies that G_j is a finite set. Then $\eta \equiv \max(G_j \cap (-\infty, 0))$ is well defined and $\eta < 0$. Assumption B.1(b) implies that there is a positive constant C such that $G_{j^*} \subseteq [-C, C]$. Let β further satisfy $\beta_{j^*}/\beta_j < -\eta/C$. Then, for all $g \in G$ such that $g_j < 0$, we have

$$\beta' g = \beta_j g_j + \beta_{j^*} g_{j^*} \leq \beta_j \eta + \beta_{j^*} C < 0. \tag{B.1}$$

[18] We were not able to obtain an analogous result in the more general multinomial choice case because (i) cycles longer than 2 would need to be considered, and (ii) the simultaneous variation of X^k_{it} for all k would also need to be taken into account.

Consider $\tilde{G} = \{g \in G : \beta' g > 0\}$. Then (B.1) implies that for all $g \in G$ such that $g_j < 0$, we have $g \notin \tilde{G}$. That implies that $\text{coni}(\tilde{G})$ contains no point whose jth element is negative. The proof of Theorem 3.1 shows that $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in \tilde{G}\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$ under Assumptions 3.1 and 3.2, which implies that $\text{coni}(\tilde{G}) = \text{coni}(\mathcal{G})$. Thus, $cc(\mathcal{G})$ also contains no point whose jth element is negative. Let b^* be the same as β except that $b^*_j > \beta_j$. Let $b = b^*/\|b^*\|$. Then $(b^*)'g \geq \beta' g$ for all $g \in \text{coni}(\mathcal{G})$. Because $\beta' g \geq 0$ for all $g \in \mathcal{G}$, we have $(b^*)'g \geq 0$ for all $g \in \mathcal{G}$, and thus $b'g \geq 0$ for all $g \in \mathcal{G}$. This implies that $Q(b) = 0$. □

C Appendix: Monte Carlo Results for Instrumental Function-Based Estimator

In this section we report the Khan and Tamer (2009)-variant of the moment inequality estimator. In this approach, rather than estimating the conditional choice probabilities and plugging them into the CM inequalities, we transform the conditional moment inequalities into unconditional moment inequalities for estimation.

The instrumental functions are indicator functions of hypercubes in the space of $\mathbf{X}_i$, where $\mathbf{X}_i = (\text{vec}(\mathbf{X}_{i1})', \text{vec}(\mathbf{X}_{i2})')'$. There are many ways to choose and weight the hypercubes to use, among which Khan and Tamer (2009) suggests to use the hypercubes formed by pairs of observations in the data. That suggests the criterion function below:

$$Q_n^{IF}(b) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} [\bar{g}_n(\mathbf{X}_i, \mathbf{X}_j)' b]_-,$$

where

$$\bar{g}_n(\underline{\mathbf{x}}, \bar{\mathbf{x}}) = n^{-1} \sum_{i=1}^{n} g_i(\underline{\mathbf{x}}, \bar{\mathbf{x}})$$

$$g_i(\underline{\mathbf{x}}, \bar{\mathbf{x}}) = \Delta \mathbf{X}_i \Delta \mathbf{Y}_i 1\{\underline{\mathbf{x}} \leq \mathbf{X}_i < \bar{\mathbf{x}}\}.$$

When implementing this approach, we were faced with two problems: (1) there are too many pairs involved for our sample sizes (e.g. 499,500 pairs for $n = 1000$), which makes computation very difficult, and (2) most of the hypercubes end up being empty simply due to the the fact that our $\mathbf{X}_i$ is 12 dimensional (3 variables $\times$ 2 time periods $\times$ 2 inside alternatives), which means that the criterion function often does not give a meaningful estimate.

For those reasons, we use the high-dimensional version of the hypercubes suggested in Andrews and Shi (2013) instead. In our design our variables are supported in the unit interval. Thus, we first evenly divide $[0, 1]$ into q subintervals ($q = 3$ for $n = 250$, 4 for $n = 500$, 5 for $n = 1000$, and 6 for $n = 2000$). Then use all the hypercubes that are the Cartesian products of two such sub-intervals and ten copies of $[0, 1]$. Let the collection of all such hypercubes be denoted by $\mathcal{C}$. Specifically, we form

$$Q_n^{IF}(b) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{C \in \mathcal{C}} [\bar{g}_n(C)' b]_-,$$

where $\bar{g}_n(C) = n^{-1} \sum_{i=1}^{n} g_i(C)$ and $g_i(C) = \Delta X_i \Delta Y_i 1\{X_i \in C\}$. We do not divide up all dimensions of the space of $\mathbf{X}_i$ precisely to avoid the same difficulties that arise with the Khan and Tamer (2009) instrumental functions discussed above.

We take the Cauchy design in Section 4.1 and report statistics of the instrumental function based estimator of β_2 in Table 7 below. Comparing to Table 2, we can see that the instrumental function-based CM estimator has larger bias and standard deviation. In addition, the standard deviation decreases slower with the sample size. For this reason, we focus on the estimator based on the nonparametric estimator of $\mathbf{p}(\cdot,\cdot)$ in the main text.

Table 7: Monte Carlo Performance of Estimators of β_2 (Cauchy Design, $\beta_{0,2} = 0.5$)

n	BIAS	SD	rMSE	25% quantile	median	75% quantile
Instrumental Function-Based CM Estimator						
500	-0.1132	0.1496	0.1876	0.2847	0.3831	0.4830
1000	-0.0808	0.1172	0.1424	0.3402	0.4162	0.4955
2000	-0.0471	0.0970	0.1078	0.3853	0.4501	0.5151