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

(**Renovation in Progress. Please excuse the debris.**)

April 17, 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 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.

1 Introduction

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

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

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,

*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 Johns Hopkins, Northwestern, NYU, UC Riverside, UNC, UT-Austin, the 2016 Seattle-Vancouver Econometrics Conference and the 2015 Xiamen/WISE Econometrics Conference in Honor of Takeshi Amemiya for useful comments. 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.

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, normalize $\|\beta\| = 1$, $X_{it}^0 = \mathbf{0}_{d_x}$ and $A_i^0 = 0 = \epsilon_{it}^0$. We will not impose location normalization for 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. Our assumptions will rule out dynamic panel models where X_{it}^k may include lagged values of $(Y_{it}^k)_{k=1}^K$.[1]

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 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. Then we show that these moment inequalities point identify β under an intuitive necessary and sufficient condition on quantities directly identifiable from the data. Two sets of sufficient conditions for uniform point identification are subsequently provided. Notably, one of the two sets of sufficient conditions allows all regressors to be bounded. We finally propose a consistent estimator for β, the computation of which requires only convex optimization.

To our knowledge, this is the first paper that deals with the incidental parameter problem while achieving point identification for semi-parametric panel multinomial choice models.[2] For partial identification in these models, Pakes and Porter (2015) propose an alternative approach.[3] Pakes and Porter construct their inequality restrictions by ranking the options according to their conditional probabilities of being chosen. By comparison, this paper contrasts the choice probabilities of all options across time periods. While the identification inequalities obtained in both papers reduce to that of Manski 1987 in the binary choice case, they are very different in general multinomial choice models.

The literature on semi-parametric panel binary choice models is large. Manski (1987) proposed the maximum score approach for identification and estimation. Abrevaya (2000) proposes a general

[1] Honoré and Kyriazidou (2000) propose a maximum score approach for dynamic panel data binary choice models.

[2] Abrevaya (1999) proposes a maximum rank-correlation estimator for panel transformation models. His approach does not apply to discrete choice models due to a strict monotonicity requirement on the transformation function.

[3] A recent version of Pakes and Porter also considers point identification.

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. Identification conditions in these papers are non-nested with ours. Chamberlain (2010) shows the impossibility of point identification in a binary choice special case of the model described by Eqs. (1.1) and (1.2) when X_{it} is bounded and contains a time dummy. This impossibility result is implied by our necessity result (Theorem B.1) 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 one of our sufficiency results (Theorem 3.2).

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.[4]

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 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 modelling in empirical IO. This section also contains an illustrative empirical application using aggregate supermarket scanner data. Section 6 concludes.

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

2 Preliminaries

In this section, we describe the concept of cyclic monotonicity and its connection to multinomial choice models. We begin by giving 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 and only if* [5]

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

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.

Above we defined both a cyclic monotonicity concept for $R^K \to R^K$ functions, 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.

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 $\boldsymbol{\epsilon} = (\epsilon^1, \ldots, \epsilon^K)'$. Then we can define a function that is a stepping stone for applying cyclic monotonicity in the multinomial choice context. The function, which McFadden (1978, 1981) called the "social surplus function," is the expected utility obtained from the choice problem:

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

[5] 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."

The following lemma shows that this function is convex, that the gradient of it is the choice probability function, and finally that the choice probability function is cyclic mononotone. The first three parts of the lemma are already known in the literature, for example in McFadden (1981), and the last part is immediately implied by the previous parts and Proposition 1. Nonetheless, we give a self-contained proof in the appendix 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) $\mathcal{G}(\cdot)$ *is convex on* R^K*,*

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

(c) $\mathbf{p}(\mathbf{u}) = \nabla\mathcal{G}(\mathbf{u})$*, 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.[6] 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) $(\boldsymbol{\epsilon}_{i1} \sim \boldsymbol{\epsilon}_{i2})|(\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2})$*, and*

(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* $\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}$*.*

[6] See also Melo, Pogorelskiy, and Shum (2015) and Chiong and Shum (2016) for other econometric applications of cyclic monotonicity.

Remark. (i) Part (a) of the assumption is the multinomial version of the the group homogeneity assumption of Manski (1987). The assumption 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 eliminate the fixed effect. This assumption rules out dynamic panel models, but it allows $\boldsymbol{\epsilon}_{it}$ to be correlated with the covariates, and allows 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 within one time period can have arbitrary joint distribution, 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 let

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

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 monotinicity with respect to length-2 cycles, we obtain, for any η_1, η_2 and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}$,[7]

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

On the other hand, by (3.2), we have

$$\mathbf{p}(\mathbf{x}_1'\beta + \mathbf{a}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}) = E[\mathbf{Y}_{i1} | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2, \mathbf{A}_i = \mathbf{a}], \tag{3.5}$$

and by (3.3),

$$\mathbf{p}(\mathbf{x}_2'\beta + \mathbf{a}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}) = E[\mathbf{Y}_{i2} | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2, \mathbf{A}_i = \mathbf{a}], \tag{3.6}$$

Combining (3.4), (3.5), and (3.6), we have

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

Take the conditional expectation given $\mathbf{X}_{i1}, \mathbf{X}_{i2}$ of both sides, and we get,

$$(E[\mathbf{Y}_{i1}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] - E[\mathbf{Y}_{i2}' | \mathbf{X}_{i1}, \mathbf{X}_{i2}])(\mathbf{X}_{i1}'\beta - \mathbf{X}_{i2}'\beta) \geq 0 \text{ pointwise.} \tag{3.8}$$

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

[7] Longer cycles are discussed later in the paper.

β with additional assumptions as discussed below. These inequalities reduce to the rank correlation conditon in Manski (1987) when $K = 1$.[8]

We summarize the result of the derivation in a lemma below. The proof for the lemma has already been given above.

Lemma 3.1. *Under Assumption* 3.1*,*

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

The key strategy used above to integrate out $\mathbf{A}_i$ is to represent $\mathbf{p}(\beta'\mathbf{x}_1 + \mathbf{a}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a})$ and $\mathbf{p}(\beta'\mathbf{x}_2 + \mathbf{a}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a})$ by the conditional choice probabilities from two different time periods for the same individual. Thus, to effectively use the cyclic monotonicity of $\mathbf{p}(\cdot, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a})$ with respect to longer cycles, we need individuals to be observed in more than two time periods. The extension in Section 5 discusses how longer cycles can be used when longer panel is available. The next subsection shows that the length-2 cycles are often enough for point identification.

3.2 Point Identification of Model Parameters

To see the amount of identification information the inequalities in (3.8) contain, rewrite it as

$$E[\Delta\mathbf{Y}'_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}]\Delta\mathbf{X}'_i\beta \geq 0 \tag{3.10}$$

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:

$$\mathcal{G} = \text{supp}(\mathbf{g}) = \text{supp}(\Delta\mathbf{X}_i E[\Delta\mathbf{Y}_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}]). \tag{3.11}$$

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 ID.

First, we impose regularity conditions on the unobservables:

Assumption 3.2. (a) $\Pr(\text{supp}(\boldsymbol{\epsilon}_{it}|\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2}) = R^K) > 0$.

(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,*

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

[8]Note that $E[\mathbf{Y}'_{i1}|\mathbf{X}_{i1}, \mathbf{X}_{i2}]$ is K-dimensional, and for this simple reason, the relationship below that a naive generalization of Manski's (1987) maximum score approach to multinomial choice would rely on does not hold:

$$\mathbf{X}'_{i1}\beta > \mathbf{X}'_{i2}\beta \Leftrightarrow E[\mathbf{Y}_{i1}|\mathbf{X}_{i1}, \mathbf{X}_{i2}] > E[\mathbf{Y}_{i1}|\mathbf{X}_{i1}, \mathbf{X}_{i2}]. \tag{3.9}$$

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

$$\Delta \mathbf{X}_i E[\Delta \mathbf{Y}_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}] = \sum_{k=1}^{K} \Delta X_i^k E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \tag{3.12}$$

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 restrictions on the covariates which we can impose to 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). Then

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

   Consequently, $\text{supp}(\Delta X_i^k E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]) = \text{supp}(\Delta X_i^k \text{sign}(E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]))$. From this we can see that it is sufficient to impose richness assumptions on the conditional support of ΔX_i^k and that of $-\Delta X_i^k$ given $\Delta \mathbf{X}_i^{-k} = 0$ because $\text{sign}(E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}])$ is either 1 or -1. We thus define

   $$G_I \equiv \cup_k \text{supp}(\pm \Delta X_i^k | \Delta \mathbf{X}_i^{-k} = 0). \tag{3.13}$$

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). Then

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

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

   $$\text{supp}(\Delta X_i^1 E[\Delta Y_i^0 | \mathbf{X}_{i1}, \mathbf{X}_{i2}]) = \text{supp}(\Delta X_i^1 \text{sign}(E[\Delta Y_i^0 | \mathbf{X}_{i1}, \mathbf{X}_{i2}])). \tag{3.14}$$

   From this we can see that it is sufficient to impose richness assumptions on the conditional support of ΔX_i^1 and that of $-\Delta X_i^1$ given $\Delta_i^k = \Delta_i^1 \ \forall k$. We thus define

   $$G_{II} \equiv \text{supp}(\pm \Delta X_i^1 | \Delta X_i^k = \Delta X_i^1 \ \forall k). \tag{3.15}$$

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

$$G \equiv G_I \cup G_{II}. \tag{3.16}$$

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 jist 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.[9]

We next present an alternative to Assumption 3.3 that allows discrete covariates generally, but requires one regressor with large support.[10] Let g_{-1} be g with the first element removed. Let $G_{-1} = \{g_{-1} : \exists g_1 \in R \; s.t. (g_1, g_{-1}')' \in G\}$. Let $G_1(g_{-1}) = \{g_1 \in R : (g_1, g_{-1}')' \in G\}$. Let g_{-j}, G_{-j}, $G_j(g_{-j})$ be defined analogously for $j = 2, \ldots, d_x$.

Assumption 3.4. (a) *$G_j(g_{-j}) = \mathbb{R}$ for all g_{-j} in a subset G_{-j}^0 of G_{-j},*

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

(c) *the jth element of β, denoted by β_j, is nonzero.*

The identification result is stated using the following criterion function:

$$Q(b) = E|\min(0, E[\Delta \mathbf{Y}_i' | \mathbf{X}_{i1}, \mathbf{X}_{i2}] \Delta \mathbf{X}_i' b)|. \tag{3.17}$$

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 $b \in \{b \in R^{d_x} : \|b\| = 1\}$ and $b \neq \beta$.*

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

3.3 Examples: Continuous covariates

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, given $\Delta X_i^1 = 0$,

$$\text{supp}(\Delta X_i^2) = [-1,1]^2.$$

Obviously, $[-1,1]^2$ contains an open neighborhood of the origin; thus, Assumption 3.3 is satisfied.

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

[10] 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 (2001), our Theorem B.1).

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 satisifies Assumption 3.3.

Example 3. Suppose that the covarites 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.

3.4 Example: Discrete Covariates

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 the special j in Assumption 3.4 be 2, 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 since $\beta_2 \neq 0$.

3.5 Remarks

Special case: 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[11]

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

[11] 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 part (b) in 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 (3.17). Specifically, we obtain a consistent estimator of β as $\widehat{\beta} = \widetilde{\beta}/\|\widetilde{\beta}\|$, where

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

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

where $\Delta \hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2}) = \hat{\mathbf{p}}_2(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \hat{\mathbf{p}}_1(\mathbf{X}_{i1}, \mathbf{X}_{i2})$. 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. [12]

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

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 conditional moment inequalities because the identifying

[12] 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.

conditions in (3.8) are 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 even when our point identification assumptions do not hold.

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 $[0,1]$. Let $A_i^k = (\omega_i^k + \sum_{j=1}^3 X_{j,i1}^k)/4$ for $k = 1, 2$, $t = 1, 2$, where ω_i^k is uniform in $[0,1]$, independent across k and independent of other model primitives. Let

$$(u_{it}^0, u_{it}^1, u_{it}^2) \sim N\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0 & 0\\0 & 1 & 0.5\\0 & 0.5 & 1\end{pmatrix}\right), \tag{4.3}$$

and let $\epsilon_{it}^k = A_i^k(u_{it}^k - u_{it}^0)$, for $t = 1, 2$. Let $(u_{i1}^0, u_{i1}^1, u_{i1}^2)$ be independent of $(u_{i2}^0, u_{i2}^1, u_{i2}^2)$. Let the true coefficient parameter $\beta = (1, 0.5, 0)$. Note that for this test model, only our estimator yields consistent point estimates: Pakes and Porter (2013) only consider partial identification, and Chamberlain's (1980) conditional logit model requires the errors to be i.i.d. extreme-value distributed.

We compute the bias, standard deviation (SD) and the root mean-squared error (rMSE) of each element of $\widehat{\beta}$ defined in the previous section. The nonparametric conditional choice probabilities are estimated using the k-nearest neighbor estimator where the tuning parameter k is selected via leave-one-out cross-validation. We consider four sample sizes 250, 500, 1000, and 2000, and use 6000 Monte Carlo repetitions. The results are reported in Table 1. As we can see, the standard deviation decreases with the sample size for every element of the parameter, which is the general pattern for the bias as well.

Table 1: Monte Carlo Results (6000 Repetitions)

n	$\widehat{\beta}_1$			$\widehat{\beta}_2$			$\widehat{\beta}_3$		
	BIAS	SD	rMSE	BIAS	SD	rMSE	BIAS	SD	rMSE
250	.0158	.0694	.0712	-.0890	.1375	.1638	-.0173	.1384	.1394
500	.0199	.0396	.0444	-.0682	.0918	.1143	-.0128	.1009	.1017
1000	.0183	.0288	.0341	-.0525	.0664	.0846	-.0149	.0750	.0764
2000	.0163	.0216	.0270	-.0412	.0483	.0635	-.0147	.0527	.0547

5 Multiple Periods

We have thus far focused on two-period panel data sets for ease of exposition. Our method naturally generalizes to mutlipe-period panel data sets as well. Suppose that there are T time periods. Then one can use cycles with length up to T to form the moment inequalities. To begin, consider $t_1, t_2, \ldots, t_T \in \{1, 2, \ldots, T\}$, 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[13]

$$\sum_{m=1}^{T} (\mathbf{X}_{it_m} - \mathbf{X}_{it_{m+1}})' E[\mathbf{Y}_{it_m} | \mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_T}] \geq 0. \tag{5.1}$$

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

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

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

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 $\hat{\mathbf{p}}(\mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_T})$ can be noisy when $t_1, \ldots, t_T$ 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, where the estimator is defined just as above except that $t_1, \ldots, t_T$ is replaced by t_1, t_2, and $\sum_{m=1}^{T}$ is replaced by $\sum_{m=1}^{2}$.

Remark. It might be possible to obtain weaker conditions for point identification when longer cycles are used, but we were not able to write down a clean set of conditions for that.

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. Such models are often encountered in empirical industrial organization.[14] 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

[13] Note that when $t_1, \ldots, t_T$ are all distinct from each other, these inequalities come from the cyclic monotonicity with respect to length T cycles. Otherwise, they come from the cyclic monotonicity with respect to cycles of shorter length. By considering all positive combinations $\{t_1, \ldots, t_T\}$ the above expression represents inequalities resulted from cyclic monotonicity with respect to all cycles of length up to T.

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

level for each choice option. To be precise, we observe $(\boldsymbol{S}_{ct}, \mathbf{X}_{ct} = (X_{ct}^{1,\prime}, \ldots, X_{ct}^{K,\prime})')_{c=1}^{n}{}_{t=1}^{T}$ 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.

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:

(a) The error term $(\boldsymbol{\epsilon}_{ic1} \sim \boldsymbol{\epsilon}_{ic2} \sim \cdots \sim \boldsymbol{\epsilon}_{icT})|\eta_c, \mathbf{A}_{ic}$, where η_c is a random variable of arbitrary dimension that captures all market level variation including the variation of $(\mathbf{X}_{c1}, \ldots, \mathbf{X}_{cT})$ across c, 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 that, for any cycle $t_1, t_2, \ldots, t_M, t_{M+1} = t_1$ in $\{1, \ldots, T\}$,

$$\sum_{m=1}^{M} E(\mathbf{Y}_{ict_m}'|\eta_c)(\mathbf{X}_{c,t_m}'\beta - \mathbf{X}_{c,t_{m+1}}'\beta) \geq 0, \ a.s. \tag{6.2}$$

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 $\boldsymbol{S}_{ct}$.[15] To avoid nonparametric estimation of the conditional choice probability is the reason that we introduce the common shock η_c here.

Now, we can construct a consistent estimator of β. For simplicity, consider length-2 cycles only in the estimator. The estimator is defined as $\widehat{\beta} = \widetilde{\beta}/\|\widetilde{\beta}\|$, where

$$\widetilde{\beta} = \arg \min_{b \in R^{d_x} : \max_{j=1,\ldots,J} |b_j| = 1} Q_n(b), \quad \text{and} \tag{6.3}$$

$$Q_n(b) = n^{-1} \sum_{c=1}^{n} \sum_{1 \leq s < t \leq T} \left[(b'\mathbf{X}_{cs} - b'\mathbf{X}_{ct})(\boldsymbol{S}_{cs} - \boldsymbol{S}_{ct}) \right]_- . \tag{6.4}$$

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

[15] If $\inf_{c,t} n_{ct}$ grows fast enough with $C \times T$, 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 $\log(C \times T)/\min_{c,t} n_{ct} \to 0$.

6.1 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.[16] 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 2).

Table 2: 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.

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.[17] Since any price discounts during a sale will be captured in the 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 3.

The point estimates are reported in Table 4. One interesting observation from the table is that the sign of the interaction term is negative, indicating that consumers are more price sensitive when

[16] 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.

[17] 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 identification condition, Assumption 3.3, hold.

Table 3: Summary Statistics

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

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).

Table 4: 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

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. In ongoing work, we are considering the possible extension of these ideas to other models and economic settings.

Throughout this paper, we have focused on estimation under the assumption that the conditions for point identification are satisfied. In the case that these conditions are not satisfied, the parameters will only be partially identfied, and we might consider an alternative inferential approach for this case based on Freyberger and Horowitz (2013) or based on the general methods for conditional moment inequalities, for example, Chernozhukov, Lee, and Rosen (2013) and Andrews and Shi (2013). Since this approach is quite different in spirit to the methods described so far, we do not discuss it here.

References

[1] J. Abrevaya. Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics*, 93: 203-228, 1999.

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

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

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

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

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

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

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

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

[12] A. Chesher. Instrumental Variables. *Journal of Econometrics*, 139:15-34, 2007.

[13] K. Chiong and M. Shum (2016). Random Projection Estimation of Discrete-Choice Models with Large Choice Sets. Working paper, California Institute of Technology.

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

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

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

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

[18] —. Large Sample Properties of the Maximum Rank Correlation Estimator in Generalized Regression Models. Working paper, 1988.

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

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

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

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

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

[24] G. W. Imbens. Nonparametric Estimation of Average Treatment Effect Under Exogeneity: A Review. *Review of Economics and Statistics* 86(1):1-29, 2004.

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

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

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

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

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

[30] C. F. Manski. Identification of Binary Response Models. *JASA*, 83:729-738, 1988.

[31] D. L. McFadden. Modelling 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] E. Melo, K. Pogorelskiy, and M. Shum (2015). Testing the quantal response hypothesis. Working paper, California Institute of Technology.

[34] 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.

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

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

[37] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[38] P. Rosenbaum and D. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 70:41-55, 1983.

[39] J. H. Stock and M. W. Watson. *Introduction to Econometrics*. 3rd ed. Pearson Publishing, 2010.

[40] 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

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

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 $\lim_{\eta\to 0}[\mathcal{G}(u_*^1, \ldots, u_*^K + \eta) - \mathcal{G}(u_*^1, \ldots, u_*^K)]/\eta$ exists. We show this using the bounded convergence theorem. First observe that

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

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. It is clear from the definition of $Q(\cdot)$ that $Q(b) \geq 0$. Also, by Equation (3.10), we have $Q(\beta) = 0$.

We now 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 show that (i) $\{g \in R^{d_x} : \beta' g > 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \tilde{G}\}$ and (ii) $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in \tilde{G}\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. Both (i) and (ii) together immediately imply that $\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. By definition of coni$(\cdot)$ above,

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

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, $g \in \tilde{G}$. Because g_0 is taken arbitrarily, this shows result (i).

To show (ii), consider an arbitrary point $\tilde{g} \in cone(\tilde{G})$. Then there exists $\lambda \geq 0$ and $g \in G$ such that $\beta' g > 0$ and $\tilde{g} = \lambda g$. 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

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

This and (A.16) together imply that $g \in \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. Since $\tilde{g} = \lambda g$ for some $\lambda \geq 0$, we have $\tilde{g} \in \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. Because $\tilde{g}$ is an arbitrary point in $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in \tilde{G}\}$, this shows result (ii).

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

$$\begin{aligned}
&E[Y_{i2}^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = 0, X_{i1}^k = x_\dagger^k] \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i2}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i2}^{k'} + A_i^{k'} + \epsilon_{i2}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i2}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= \Pr\left(\beta' x_*^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i1}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&> \Pr\left(\beta' x_\dagger^k + A_i^k + \epsilon_{i1}^k \geq \max_{k'=0,\ldots,K: k' \neq k} \beta' X_{i1}^{k'} + A_i^{k'} + \epsilon_{i1}^{k'} \middle| X_{i2}^k = x_*^k, \Delta \mathbf{X}_{it}^{-k} = 0, X_{i1}^k = x_\dagger^k \right) \\
&= E[Y_{i1}^k | X_{i2}^k = x_*^k, \Delta \mathbf{X}_i^{-k} = 0, X_{i1}^k = x_\dagger^k],
\end{aligned} \tag{A.18}$$

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 $\mathcal{G}$:

$$-E[\Delta Y_i^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K] g. \tag{A.19}$$

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

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

This and (A.19) together imply that $g \in \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. Since $\tilde{g} = \lambda g$ for some $\lambda \geq 0$, we have $\tilde{g} \in \{\lambda g : \lambda \in R, \lambda \geq 0, g \in \mathcal{G}\}$. Because $\tilde{g}$ is an arbitrary point in $\{\lambda g : \lambda \in R, \lambda \geq 0, g \in \tilde{G}\}$, this shows result (ii).

Inequality (A.20) follows from the derivation:

$$\begin{aligned}
&E[Y_{i2}^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K] \\
&= \Pr\left(\max_{k=1,\ldots,K} \beta' x_*^k + A_i^k + \epsilon_{i2}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&= \Pr\left(\max_{k=1,\ldots,K} \beta' x_*^k + A_i^k + \epsilon_{i1}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&< \Pr\left(\max_{k=1,\ldots,K} \beta' x_\dagger^k + A_i^k + \epsilon_{i1}^k \leq 0 \,\middle|\, X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k \right) \\
&= E[Y_{i1}^0 | X_{i2}^k = x_*^k, X_{i1}^k = x_\dagger^k \ \forall k = 1, \ldots, K],
\end{aligned} \tag{A.21}$$

where the arguments for each steps are the same as those for the corresponding steps in (A.18).

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

$$\text{coni}(\tilde{G}) \subseteq \text{coni}(\mathcal{G}). \tag{A.22}$$

Below we show that

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

This combined with (A.22) proves (A.14) and thus proves the theorem.

Now we show (A.23). 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

$$\tilde{G}^0 \subseteq \tilde{G}. \tag{A.24}$$

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

$$d := g_{0,j} + g_{0,-j}' \beta_{-j}/\beta_j > 0. \tag{A.25}$$

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 jth element is $g_{m,j}$ and the rest of whose elements form $g_{m,-j}$, for $m = 1, \ldots, M$. Then $g_m \in \tilde{G}^0$ for $m = 1, \ldots, M$ because $g_{m,-j} \in G^0_{-j}$ by construction and $d_{m,j} > -g'_{m,-j}\beta_{-j}/\beta_j$ 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,

$$g_0 \in \text{coni}(\tilde{G}). \tag{A.26}$$

Therefore, $\{g \in R^{d_x} : \beta' g \geq 0\} \subseteq \text{coni}(\tilde{G})$. □

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

$$\widetilde{\beta} \to_p \beta/\|\beta\|_\infty. \tag{A.27}$$

This implies that $\widehat{\beta} \to_p \beta$ because $\widehat{\beta} = \widetilde{\beta}/\|\widetilde{\beta}\|$ 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.27). Let

$$Q(b) = E\left[b'(\mathbf{X}_{i1} - \mathbf{X}_{i2})\left(\mathbf{p}_1(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \mathbf{p}_2(\mathbf{X}_{i1}, \mathbf{X}_{i2})\right)\right]_-. \tag{A.28}$$

Under Assumption 3.1, the identifying inequalities (3.8) hold, which implies that

$$Q(\beta) = Q(\beta/\|\beta\|_\infty)) = 0. \tag{A.29}$$

Assumption 4.1 and Theorem 3.1 together imply that, for any $b \neq \beta/\|\beta\|_\infty$ such that $\|b\|_\infty = 1$,

$$\Pr\left(b'(\mathbf{X}_{i1} - \mathbf{X}_{i2})\left(\mathbf{p}_1(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \mathbf{p}_2(\mathbf{X}_{i1}, \mathbf{X}_{i2})\right) < 0\right) > 0. \tag{A.30}$$

Thus, for any $b \neq \beta/\|\beta\|_\infty$ such that $\|b\|_\infty = 1$, we have that $Q(b) > 0$. 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,

$$\inf_{b \in R^{d_x} : \|b\|_\infty = 1, \|b - \beta\| > \varepsilon} Q(b) \geq \delta. \tag{A.31}$$

If in addition, we can show the uniform convergence of $Q_n(b)$ to $Q(b)$, then the consistency of $\widehat{\beta}$ follows from standard consistency arguments (see, e.g., Newey and McFadden (1994)).

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

$$\sup_{b \in R^{d_x} : \|b\|_\infty = 1} |Q(b) - Q_n(b)| \to_p 0. \tag{A.32}$$

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

$$|Q_n(b) - Q_n(b^*)|$$

$$
\begin{aligned}
&\leq n^{-1}\sum_{i=1}^{n}\left|(b-b^*)'(\mathbf{X}_{i1}-\mathbf{X}_{i2})\left(\hat{\mathbf{p}}_1(\mathbf{X}_{i1},\mathbf{X}_{i2})-\hat{\mathbf{p}}_2(\mathbf{X}_{i1},\mathbf{X}_{i2})\right)\right| \\
&\leq n^{-1}\sum_{i=1}^{n}\|b-b^*\|\|(\mathbf{X}_{i1}-\mathbf{X}_{i2})\left(\hat{\mathbf{p}}_1(\mathbf{X}_{i1},\mathbf{X}_{i2})-\hat{\mathbf{p}}_2(\mathbf{X}_{i1},\mathbf{X}_{i2})\right)\| \\
&\leq 2n^{-1}\sum_{i=1}^{n}\|\mathbf{X}_{i1}-\mathbf{X}_{i2}\|\|b-b^*\|. \qquad \text{(A.33)}
\end{aligned}
$$

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

$$
\begin{aligned}
&\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\Pr\left(\sup_{b,b^*\in R^{d_x},\|b\|_\infty=\|b^*\|_\infty=1,\|b-b^*\|\leq\delta}|Q_n(b)-Q_n(b^*)|>\varepsilon\right) \\
&\leq \lim_{\delta\downarrow 0}\limsup_{n\to\infty}\Pr\left(2\delta n^{-1}\sum_{i=1}^{n}\|\mathbf{X}_{i1}-\mathbf{X}_{i2}\|>\varepsilon\right) \\
&\leq \lim_{\delta\downarrow 0}\limsup_{n\to\infty}\Pr\left(2n^{-1}\sum_{i=1}^{n}\|\mathbf{X}_{i1}-\mathbf{X}_{i2}\|>\varepsilon/\delta\right) \\
&= 0, \qquad \text{(A.34)}
\end{aligned}
$$

where the first inequality holds by (A.33) and the equality holds by Assumption 4.1(c). 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.32), it suffices to show that for all $b \in R^{d_x} : \|b\|_\infty = 1$, we have

$$
Q_n(b) \to_p Q(b). \qquad \text{(A.35)}
$$

For this purpose, let

$$
\widetilde{Q}_n(b) = n^{-1}\sum_{i=1}^{n}\left[(b'\mathbf{X}_{i1}-b'\mathbf{X}_{i2})(\mathbf{p}_1(\mathbf{X}_{i1},\mathbf{X}_{i2})-\mathbf{p}_2(\mathbf{X}_{i1},\mathbf{X}_{i2}))\right]_-. \qquad \text{(A.36)}
$$

By Assumption 4.1(c) 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:

$$
\begin{aligned}
&|\widetilde{Q}_n(b)-Q_n(b)| \\
&\leq n^{-1}\sum_{i=1}^{n}\sum_{j=1,2}\left|(b'\mathbf{X}_{i1}-b'\mathbf{X}_{i2})(\hat{\mathbf{p}}_j(\mathbf{X}_{i1},\mathbf{X}_{i2})-\mathbf{p}_j(\mathbf{X}_{i1},\mathbf{X}_{i2}))\right|, \\
&\leq 2\sup_{\mathbf{x}_1,\mathbf{x}_2\in\boldsymbol{\mathcal{X}}}\sup_{j=1,2}\|\hat{\mathbf{p}}_j(\mathbf{x}_1,\mathbf{x}_2)-\mathbf{p}_j(\mathbf{x}_1,\mathbf{x}_2)\|n^{-1}\sum_{i=1}^{n}\|b'\mathbf{X}_{i1}-b'\mathbf{X}_{i2}\|, \\
&\to_p 0, \qquad \text{(A.37)}
\end{aligned}
$$

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

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 $b \in \{b \in R^{d_x} : \|b\| = 1\}$ such that $b \neq \beta$.*

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

[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_{it}^k for all k would also need to be taken into account.

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

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

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 $Q(b) = 0$. □