ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 86, No. 2 (March, 2018), 737–761

ESTIMATING SEMI-PARAMETRIC PANEL MULTINOMIAL CHOICE MODELS USING CYCLIC MONOTONICITY

XIAOXIA SHI
Department of Economics, University of Wisconsin-Madison

MATTHEW SHUM
Division of Humanities and Social Sciences, Caltech

WEI SONG
Gregory and Paula Chow Center for Economic Research (GCCER) and Department of Statistics, School of Economics, Xiamen University

Econometrica, Vol. 86, No. 2 (March, 2018), 737–761

ESTIMATING SEMI-PARAMETRIC PANEL MULTINOMIAL CHOICE MODELS USING CYCLIC MONOTONICITY

XIAOXIA SHI
Department of Economics, University of Wisconsin-Madison

MATTHEW SHUM
Division of Humanities and Social Sciences, Caltech

WEI SONG
Gregory and Paula Chow Center for Economic Research (GCCER) and Department of Statistics, School of Economics, Xiamen University

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.

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, A_i^k is an agent/alternative-specific fixed effect, 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 semi-parametric settings, we

Xiaoxia Shi: xshi@ssc.wisc.edu
Matthew Shum: mshum@caltech.edu
Wei Song: wsong@xmu.edu.cn

We thank Khai Chiong, Federico Echenique, Bruce E. Hansen, Whitney Newey, Jack R. Porter, and seminar audiences at Brown, Michigan State, Ohio State, Chicago Booth, Johns Hopkins, Northwestern, NYU, UC Riverside, UNC, Princeton, Cornell, the 2016 Seattle-Vancouver Econometrics Conference, the 2015 Xiamen/WISE Econometrics Conference in Honor of Takeshi Amemiya, the 2016 Interaction Conference at Northwestern, the 2016 Conference on Nonstandard Problems at Duke, and the 2017 Cornell-Penn State IO and Econometrics Conference 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 for the year 2014–2015.

 https://doi.org/10.3982/ECTA14115

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 (2013) proposed 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) proposed 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) used a maximum score-type estimator for a dynamic panel binary choice model. Abrevaya (2000) proposed 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) generalized 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) based 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 $E[Y_i^k | X_i^1, \ldots, X_i^K]$ 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 assumed that the error terms are i.i.d. across k, while Fox relaxed the i.i.d. assumption to exchangeability. Exchangeability (or the rank-order property) is not used in our approach. Lewbel (2000) considered identification using a special regressor. In addition, Powell and Ruud (2008) and Ahn, Ichimura, Powell, and Ruud (2018) considered an approach based on matching individuals with equal conditional choice probabilities, which requires that the rank of a certain matrix formed from the data be deficient by exactly 1. This approach does not obviously extend to the panel data setting with fixed effects.

The existing literature 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 a generic multinomial choice model. Subsequently, in Section 3, we present the moment inequalities emerging from cyclic monotonicity for panel multinomial choice models, and give assumptions under which these inequalities suffice to point identify the parameters of interest. This section also contains examples where the point identification assumptions are verified. Section 4 presents an estimator, shows its consistency, and evaluates its performance using Monte Carlo experiments. Sections 3 and 4 focus on 2-period panels and length-2 cycles. In Section 5, we extend the discussion to longer panels and longer cycles. In Section 6, we discuss the closely related aggregate panel multinomial choice model, which is a workhorse model for demand modeling in empirical industrial organization. This section also contains an illustrative empirical application using aggregate supermarket scanner data. Section 7 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, Chapter 24), Villani (2003, Section 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

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

[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."

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:

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

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 (e.g., 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 $\mathbf{U}_{it} \equiv (U^1_{it}, \ldots, U^K_{it})'$, $\boldsymbol{\epsilon}_{it} \equiv (\epsilon^1_{it}, \ldots, \epsilon^K_{it})'$, and $\mathbf{Y}_{it} \equiv (Y^1_{it}, \ldots, Y^K_{it})'$. Let there be an observable d_x-dimensional covariate X^k_{it} for each choice k, and let U^k_{it} be a linear index of X^k_{it} plus an unobservable individual effect A^k_i:

$$U^k_{it} = \beta' X^k_{it} + A^k_i, \tag{3.1}$$

where β is a d_x-dimensional unknown parameter. Let $\mathbf{X}_{it} = (X^1_{it}, \ldots, X^K_{it})$ and $\mathbf{A}_i = (A^1_i, \ldots, A^K_i)'$. 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 group homogeneity assumption of Manski (1987), and is also imposed in Pakes and Porter (2013). 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 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 in a given period can be 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

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

Assumption 3.1(a) implies that

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

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 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}$:

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

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

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

Combining (3.4) and (3.5), we have

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

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

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

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,*

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

The extension in Section 5 discusses how longer cycles can be used when more time periods are available in the data set. 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 identification information contained in the inequalities in (3.7), we rewrite them as

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

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

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

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* $(\boldsymbol{\epsilon}_{it}|\mathbf{A}_i, \mathbf{X}_{i1}, \mathbf{X}_{i2})$ *is* R^K *with positive probability.*

(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,*

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

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

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

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 $\Delta \mathbf{X}_i^{-k} = 0$ (i.e., individual i's covariates are constant across both periods, for all choices except the kth choice), we have

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

Note that $\operatorname{supp}(\Delta X_i^k E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]) = \operatorname{supp}(\Delta X_i^k \operatorname{sign}(E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]))$ due to the fact that $E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]$ is a scalar random variable. Assumption 3.2(a) ensures that we have $\Pr(E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}] = 0 | \Delta X_i^{-k} = 0) = 0$, which implies that $\operatorname{sign}(E[\Delta Y_i^k | \mathbf{X}_{i1}, \mathbf{X}_{i2}]) \in \{-1, 1\}$ with probability 1 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 $\Delta \mathbf{X}_i^{-k} = 0$. We are thus motivated to define

$$G_{\mathrm{I}} \equiv \bigcup_k \operatorname{supp}\big(\pm \Delta X_i^k | \Delta \mathbf{X}_i^{-k} = 0\big), \tag{3.11}$$

where the conditional support of $\pm \Delta X_i^k$ is the union of the conditional support of ΔX_i^k and that of $-\Delta X_i^k$.

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

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

where $\Delta Y_i^0 = -\sum_{k=1}^K \Delta Y_i^k$. Similar arguments as above show that it is sufficient to assume a rich support for ΔX_i^1 and $-\Delta X_i^1$, which motivates us to define

$$G_{\mathrm{II}} \equiv \operatorname{supp}\big(\pm \Delta X_i^1 | \Delta X_i^k = \Delta X_i^1 \ \forall k\big). \tag{3.12}$$

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

$$G \equiv G_{\mathrm{I}} \cup G_{\mathrm{II}}. \tag{3.13}$$

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.

Our second sufficient condition allows discrete covariates generally, but requires one regressor with large support. Let g_{-1} denote g with the first element removed, and define $G_{-1} = \{g_{-1} : \exists g_1 \in R \text{ s.t. } (g_1, g'_{-1})' \in G\}$. 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) $G_{j^*}(g_{-j^*}) = R$ *for all* g_{-j^*} *in a subset* $G^0_{-j^*}$ *of* G_{-j^*},
(b) $G^0_{-j^*}$ *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:

$$Q(b) = E\big|\min\big(0, E\big[\Delta\mathbf{Y}_i'|\mathbf{X}_{i1}, \mathbf{X}_{i2}\big]\Delta\mathbf{X}_i'b\big)\big|. \tag{3.14}$$

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* $\|b\| = 1$.

REMARK: In the binary choice case, our condition 3.3 reduces to conditions similar to those in Hoderlein and White (2012), and our condition 3.4 reduces to conditions similar to those in Manski (1987). Hoderlein and White (2012) and Manski (1987) arrived at their respective conditions via distinct and mutually incompatible arguments. We arrive at both types of point identification conditions in a single framework, and by doing so were able to demonstrate the necessity of the large support condition in the presence of a discrete regressor in the binary choice case. This is reported as Theorem C.1 in Section C of the Supplemental Material (Shi, Shum, and Song (2018)), which extends, in a small way, Theorem 1 of Chamberlain (2010) to the case with general discrete regressors rather than a time dummy, and provides an alternative proof for the case of nonparametric error.[3]

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 ($d_x = 2$).

EXAMPLE 1: $\text{supp}((X^k_{it})_{t=1,2;k=1,2}) = [0,1]^8$. Then, $\text{supp}((\Delta X^k_i)_{k=1,2}) = [-1,1]^4$. Then, $G_{\text{I}} \supseteq \text{supp}(\Delta X^2_i | \Delta X^1_i = 0) = [-1,1]^2$. 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: $X^k_{it} = X_{it}$ for $k = 1, 2$, and $\text{supp}((X_{it})_{t=1,2}) = [0,1]^4$. Thus, $G_{\text{II}} = \text{supp}(\Delta X_i) = [-1,1]^2$, 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 $\text{supp}((X^1_{it})_{t=1,2}) = [0,1]^4$, $\text{supp}((X^2_{it})_{t=1,2}) = \{0,1\}^4$, and the joint support is the Cartesian product. Then, $\text{supp}(\Delta X^1_i | \Delta X^2_i = \mathbf{0}) = [-1,1]^2$. Thus, Assumption 3.3 is satisfied.

[3]Ahn et al. (2018), who studied a general model that encompasses the cross-sectional multinomial choice model, gave a point identification condition that allows discrete regressors without explicitly requiring large support. Their condition is imposed on a non-primitive quantity that depends on β. This condition may hold for a subset of values of β, but not for all values of β in $\{b \in R^K : \|b\| = 1\}$, as implied by Theorem 1 of Chamberlain (2010).

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,

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

Hence, $G \supseteq G_{\text{II}} = \{-1, 1\} \times \mathbb{R}$. 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 cyclic monotonicity (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

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

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),

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

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

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

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

In this section, we propose a computationally easy consistent estimator for β, based on Theorem 3.1. The consistency is obtained when $n \to \infty$ with T fixed. In particular, we focus on $T = 2$ and only discuss longer panels in Section 5 below. 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 (3.14). Specifically, we obtain a consistent estimator of β as $\widehat{\beta} = \overline{\beta}/\|\overline{\beta}\|$, where

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

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

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.[4,5]

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

ASSUMPTION 4.1: (a) $\max_{i=1,\ldots,n} \|\Delta\hat{\mathbf{p}}(\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \Delta\mathbf{p}(\mathbf{X}_{i1}, \mathbf{X}_{i2})\| \to_p 0$ *as* $n \to \infty$, *where* $\Delta\mathbf{p}(\mathbf{x}_1, \mathbf{x}_2) = E[\mathbf{Y}_{i2} - \mathbf{Y}_{i1} | \mathbf{X}_{i1} = \mathbf{x}_1, \mathbf{X}_{i2} = \mathbf{x}_2]$, *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:

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

The consistency result in Theorem 4.1 relies on an estimator of $\Delta\mathbf{p}(\mathbf{x}_1, \mathbf{x}_2)$ that is uniformly consistent over the observed values of $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ (i.e., Assumption 4.1). When $\text{supp}(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ is compact, any estimator that is uniformly consistent on $\text{supp}(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ satisfies this requirement. Such estimators are abundant in the nonparametric regression literature; see, for example, Devroye (1978) for nearest neighbor estimators, Hansen (2008) for kernel regression estimators, and Hirano, Imbens, and Ridder (2003) for a sieve logit estimator. When $\text{supp}(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ is not compact, at least the nearest neighbor estimators and the kernel regression estimators can still be shown to satisfy our uniform consistency requirement using the uniform convergence rate results in the references just given, provided that the observed n values of $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ are contained in a slowly expanding subset of its support with probability approaching 1. The existence of such a slowly expanding subset can be guaranteed by imposing appropriate tail probability bounds for the covariates. For example, if a kernel regression estimator is used, one can show using Theorem 2 of Hansen (2008) that it suffices to assume that the covariates have tails that are thinner than polynomial.

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, e.g., 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]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.

[5]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 Section D of the Supplemental Material.

4.1. *Monte Carlo Simulation*

Consider a trinary choice example and a 2-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 $[0, 1]$.[6] 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 $\beta_1 = 1$.

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.[7] 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. 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 I and II. 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 I), 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 setup. 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.[8]

5. LONGER PANELS ($T > 2$)

We have thus far focused on 2-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

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

[7] In Section C of the Supplemental Material, 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.

[8] 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 more slowly.

TABLE I

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

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

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

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 $\hat{\mathbf{p}}_{t_m}(\mathbf{X}_{it_1}, \ldots, \mathbf{X}_{it_L})$ 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, $L = 2$.

TABLE II

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

TABLE III

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

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 equation (5.1) with $L = T$ exhausts all the information in the random utility model and leads to the sharp identified set. We believe this is unlikely in general 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.[9] However, these properties are not straightforward to use in the panel data setting and do not lead to simple linear (in parameters) moment conditions.

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 III, 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.

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

[9]These other properties were studied in Koning and Ridder (2003).

for which we are able to derive some inference results. Such models are often encountered in empirical industrial organization.[10] 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 $(\boldsymbol{S}_{ct}, \mathbf{X}_{ct} = (X_{ct}^{1,'}, \ldots, X_{ct}^{K,'})')_{c=1,t=1}^{n,T}$ 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, that is, $\boldsymbol{S}_{ct} = I_{ct}^{-1} \sum_{i=1}^{I_{ct}} \mathbf{Y}_{ict}$.

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* $(\boldsymbol{\epsilon}_{ic1} \sim \boldsymbol{\epsilon}_{ic2})|\eta_c, \mathbf{A}_{ic}$ *and*

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

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

LEMMA 6.1: *Under Assumption 6.1, we have*

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

We no longer need to perform the nonparametric estimation of conditional choice probabilities because $E(\mathbf{Y}_{ict}|\eta_c)$ can be estimated uniformly consistently by $\boldsymbol{S}_{ct}$.[11]

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

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

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

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.

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

[11] If $\inf_{c,t} I_{ct}$ grows fast enough with $n \times T$, this estimator is uniformly consistent, that is, $\sup_c \sup_t \|\boldsymbol{S}_{ct} - E(\mathbf{Y}_{ict}|\eta_c)\| \to_p 0$. Section 3.2 of Freyberger's (2015) arguments (using Bernstein's Inequality) implies that the above convergence holds if $\log(n \times T)/\min_{c,t} I_{ct} \to 0$.

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 rather 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 coordinate 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* $\|b - \beta\| \le c_1$ *for a constant* c_1.

(e) *With probability* 1, *the conditional cumulative distribution function* $F_{\mathbf{W}_c^1|\tilde{\mathbf{W}}_c}(\cdot|\tilde{\mathbf{W}}_c)$ *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*

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

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

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. In the individual-level data setting, an analogue of Assumption 6.2(b) would hold and an $n^{-1/2}$ rate for $\widehat{\beta}$ would be obtained if the conditional choice probability were either known or estimable at a parametric rate. However, a known conditional choice probability or one estimated with parametric rate is implausible in that setting. We conjecture that with individual-level data, the noise due to estimating the conditional choice probability dominates and determines the rate. However, we have not found a way to handle this part of the noise.

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) and (e) are used to guarantee the existence of directional derivatives, while part (f) ensures that the second-order

TABLE IV

TABLE OF THE SEVEN 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)

directional derivative is bounded from below by a quadratic function.[12] The proof of the following theorem is given in Section B of the Supplemental Material.

THEOREM 6.1: *Under Assumption 6.2, we have* $\widehat{\beta} - \beta = O_p(n^{-1/2})$ *for* $\widehat{\beta}$ *defined in equation* (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.[13] 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 IV).

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 $X^k = (\text{PRICE, DEAL, PRICE*DEAL})$. 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.[14] Since any price discounts during a sale will be captured in the PRICE variable

[12]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.

[13]This data set 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.

[14]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 V
SUMMARY STATISTICS

		Min	Max	Mean	Median	Std.Dev.
10 week data	DEAL	0	1	0.4350	0	0.4749
	PRICE	0.1776	0.6200	0.3637	0.3541	0.0876
	PRICE×DEAL	0	0.6136	0.1512	0	0.1766
15 week data	DEAL	0	1	0.4488	0	0.4845
	PRICE	0.1849	0.6200	0.3650	0.3532	0.0887
	PRICE×DEAL	0	0.6091	0.1644	0	0.1888

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

The point estimates are reported in Table VI. 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 nondecreasing.

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.

APPENDIX A: PROOFS

PROOF OF LEMMA 2.1: (a) By the independence between $\mathbf{U}$ and $\boldsymbol{\epsilon}$, we have

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

TABLE VI
POINT ESTIMATES FOR BATHROOM TISSUE CHOICE MODEL

		10 Week Data	15 Week Data
β_1	DEAL	0.1053	0.0725
β_2	PRICE	−0.9720	−0.9922
β_3	PRICE×DEAL	−0.2099	−0.1017

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 $(u_*^1,\dots,u_*^K)$ denote an arbitrary fixed value of $(U^1,\dots,U^K)$, and let $u_*^0=0$. It suffices to show that $\lim_{\eta\to 0}[\mathcal{W}(u_*^1,\dots,u_*^K+\eta)-\mathcal{W}(u_*^1,\dots,u_*^K)]/\eta$ exists. We show this using the bounded convergence theorem. First observe that

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

where $\Delta(\eta,\mathbf{u}_*,\boldsymbol{\epsilon})=\max\{u_*^1+\epsilon^1,\dots,u_*^K+\eta+\epsilon^K\}-\max\{u_*^1+\epsilon^1,\dots,u_*^K+\epsilon^K\}$. Consider an arbitrary value $\boldsymbol{e}$ of $\boldsymbol{\epsilon}$ and $e^0=0$. If $e^K+u_*^K>\max_{k=0,\dots,K-1}[u_*^k+e^k]$, 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 $e^K+u_*^K<\max_{k=0,\dots,K-1}[u_*^k+e^k]$, 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,

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

By the absolute continuity of the conditional distribution of $\boldsymbol{\epsilon}$, we have $\Pr(\epsilon^K+u_*^K=\max_{k=0,\dots,K-1}[u_*^k+\epsilon^k])=0$. Therefore, almost surely,

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

Also, observe that

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

Thus, the bounded convergence theorem applies and yields

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

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

Part (d) is a direct consequence of part (c) and Proposition 1. *Q.E.D.*

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

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

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, due to $\mathcal{G}$ being the support set of $\mathbf{g}$, 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

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

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

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

By Lemma E.1 in the Supplemental Material, this implies that $\beta = b$, which contradicts the assumption that $b \neq \beta$. This concludes the proof of the lemma. *Q.E.D.*

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 $\text{coni}(\cdot)$. 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 exist 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 $\beta' g > 0$. 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 $k \in \{1, \dots, K\}$, 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

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

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

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

This and (A.16) together imply that

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

This shows result (ii).

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

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

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 that $\Delta \mathbf{X}_i^{-k} = 0$, 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 $\mathcal{G}$:

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

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

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

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

Inequality (A.21) follows from the derivation:

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

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

Under Assumption 3.4: discrete covariates. The following concepts are useful for the proof. For a set $A \subset R^K$ for an integer K, we define $-A = \{-x : x \in A\}$, and say that A is symmetric about the origin if $A = -A$.

First, we show that it is without loss of generality to assume that the set $G_{-j^*}^0$ in Assumption 3.4 is symmetric about the origin. Suppose that a set $G_{-j^*}^0$ satisfying Assumption 3.4 is not symmetric about the origin. Let $\breve{G}_{-j^*}^0 = G_{-j^*}^0 \cup (-G_{-j^*}^0)$. Then $\breve{G}_{j^*}^0$ is symmetric about the origin. We now show that for any $g_{-j^*} \in \breve{G}_{j^*}^0$, $G_{j^*}(g_{-j^*}) = R$. It suffices to consider $g_{-j^*} \in -G_{-j^*}$. Then $-g_{-j^*} \in G_{-j^*}$, which implies that $G_{j^*}(-g_{-j^*}) = R$. Note that G is defined to be the union of various conditional support sets of $\pm\Delta X_i$ each of which is symmetric about the origin. Thus, G is symmetric about the origin. That implies that $G_{j^*}(-g_{-j^*}) = -G_{j^*}(g_{-j^*})$. Therefore, $G_{j^*}(g_{-j^*}) = -R = R$, which completes the proof that $G_{j^*}(g_{-j^*}) = R$ for any $g_{-j^*} \in \breve{G}_{j^*}^0$. Last, because G is symmetric about the origin, G_{-j^*} is also symmetric about the origin, which, combined with the fact that $G_{-j^*}^0 \subseteq G_{-j^*}$, implies that $\breve{G}_{-j^*}^0 \subseteq G_{-j^*}$. Therefore, Assumption 3.4(a) remains valid if we redefine $G_{-j^*}^0$ to be $\breve{G}_{-j^*}^0$. It is straightforward to see that Assumption 3.4(b) also remains valid with the redefinition, and Assumption 3.4(c) is unaffected by the redefinition.

Thus, in the rest of the proof, we assume $G_{-j^*}^0$ to be symmetric about the origin.

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

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

Below, we show that

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

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 $\tilde{G}^0 = \{g \in R^{d_x} : g_{-j} \in G^0_{-j}, g_{j^*} > -\beta'_{-j^*} g_{-j^*}/\beta_{j^*}\}$, 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), $G^0_{-j^*}$ spans R^{d_x-1}, and is symmetric about the origin. Thus, $G^0_{-j^*}$ spans R^{d_x-1} with nonnegative weights. Then, there exists a positive integer M, weights $c_1, \ldots, c_M > 0$, and $g_{1,-j^*}, \ldots, g_{M,-j^*} \in G^0_{-j^*}$ such that $g_{0,-j^*} = \sum_{m=1}^M c_m g_{m,-j^*}$.

Let $g_{m,j^*} = (d/\sum_{m=1}^M c_m) - (g'_{m,-j^*}\beta_{-j^*}/\beta_{j^*})$ for $m = 1, \ldots, M$. Let g_m be the vector whose j^*th element is g_{m,j^*} and which 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 $g_{m,-j^*} \in G^0_{-j}$ 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, $g_0 \in \operatorname{coni}(\tilde{G}^0)$. Subsequently, by (A.25),

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

Therefore, (A.24) holds. *Q.E.D.*

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 equation (A.28). Let

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

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

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

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 constant $c > 0$, there exists a $\delta > 0$ such that

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

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, for example, Newey and McFadden (1994).

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

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

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:

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

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

$$\begin{aligned}
&\lim_{\delta\downarrow 0}\limsup_{n\to\infty} \Pr\Big(\sup_{b,b^*\in R^{d_x},\|b\|_\infty=\|b^*\|_\infty=1,\|b-b^*\|\le\delta} \left|Q_n(b) - Q_n\left(b^*\right)\right| > \varepsilon\Big) \\
&\quad\le \lim_{\delta\downarrow 0}\limsup_{n\to\infty} \Pr\left(\delta n^{-1}\sum_{i=1}^n \|\Delta\mathbf{X}_i\| > \varepsilon\right) \\
&\quad\le \lim_{\delta\downarrow 0}\limsup_{n\to\infty} \Pr\left(n^{-1}\sum_{i=1}^n \|\Delta\mathbf{X}_i\| > \varepsilon/\delta\right) \\
&\quad= 0,
\end{aligned} \tag{A.35}$$

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

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

For this purpose, let

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

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 deriva-

tion:

$$
\begin{aligned}
&\left|\widetilde{Q}_n(b) - Q_n(b)\right| \\
&\quad \le n^{-1}\sum_{i=1}^{n}\left|\left(b'\Delta\mathbf{X}_i\right)\left(\Delta\hat{\mathbf{p}}(\mathbf{X}_{i1},\mathbf{X}_{i2}) - \Delta\mathbf{p}(\mathbf{X}_{i1},\mathbf{X}_{i2})\right)\right| \\
&\quad \le \left(\sup_{i=1,\ldots,n}\left\|\Delta\hat{\mathbf{p}}_t(\mathbf{X}_{i1},\mathbf{X}_{i2}) - \Delta\mathbf{p}_t(\mathbf{X}_{i1},\mathbf{X}_{i2})\right\|\right)\left(n^{-1}\sum_{i=1}^{n}\left\|b'\mathbf{X}_{i1} - b'\mathbf{X}_{i2}\right\|\right) \\
&\quad \to_p 0,
\end{aligned}
\tag{A.38}
$$

where the convergence holds by Assumptions 4.1(a) and (b). Therefore, the theorem is proved. *Q.E.D.*

REFERENCES

ABREVAYA, J. (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics*, 95, 1–23. [738]

AHN, H., H. ICHIMURA, J. POWELL, AND P. RUUD (2018): "Simple Estimators for Invertible Index Models," *Journal of Business and Economic Statistics*, 36, 1–10. [738,744]

ANDREWS, D., AND X. SHI (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666. [746,747]

BERRY, S., AND P. HAILE (2014): "Identification in Differentiated Products Markets Using Market-Level Data," *Econometrica*, 82, 1749–1797. [750]

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 65, 841–890. [750]

BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): "Set Identified Linear Models," *Econometrica*, 80, 1129–1155. [746]

BROWNING, M. (1989): "A Nonparametric Test of the Life-Cycle Rational Expectation Hypothesis," *International Economic Review*, 30, 979–992. [738]

CHAMBERLAIN, G. (1980): "Analysis of Variance With Qualitative Data," *Review of Economic Studies*, 47, 225–238. [747]

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78, 159–168. [744]

CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737. [746]

DEVROYE, L. P. (1978): "The Uniform Convergence of Nearest Neighbor Regression Function Estimators and Their Applications in Optimization," *IEEE Transactions on Information Theory*, IT-24, 142–151. [746]

FOX, J. (2007): "Semi-Parametric Estimation of Multinomial Discrete-Choice Models Using a Subset of Choices," *RAND Journal of Economics*, 38, 1002–1029. [738]

FREYBERGER, J. (2015): "Asymptotic Theory for Differentiated Product Demand Models With Many Markets," *Journal of Econometrics*, 185, 162–181. [750]

FREYBERGER, J., AND J. HOROWITZ (2015): "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation," *Journal of Econometrics*, 189, 41–53. [746]

HAN, A. (1987): "Nonparametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316. [745]

HANSEN, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation With Dependent Data," *Econometric Theory*, 24, 726–748. [746]

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [746]

HODERLEIN, S., AND H. WHITE (2012): "Nonparametric Identification in Nonseparable Panel Data Models With Generalized Fixed Effects," *Journal of Econometrics*, 168, 300–314. [744]

HONORÉ, B., AND E. KYRIAZIDOU (2000): "Panel Discrete Choice Models With Lagged Dependent Variables," *Econometrica*, 68, 839–874. [738]

HONORÉ, B., AND A. LEWBEL (2002): "Semi-Parametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," *Econometrica*, 70, 2053–2063. [738]

HOROWITZ, J. (2009): *Semi-Parametric and Nonparametric Methods in Econometrics* (Second Ed.). Springer. [739]

KHAN, S., AND E. TAMER (2009): "Inference on Endogenously Censored Regression Models Using Conditional Moment Inequalities," *Journal of Econometrics*, 152, 104–119. [746,747]

KHAN, S., F. OUYANG, AND E. TAMER (2016): "Adaptive Rank Inference in Semiparametric Multinomial Response Models," Working Paper. [738]

KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, 191–219. [751]

KONING, R. H., AND G. RIDDER (2003): "Discrete Choice and Stochastic Utility Maximization," *Econometric Journal*, 6, 1–27. [749]

LEWBEL, A. (1998): "Semi-Parametric Latent Variable Estimation With Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121. [738]

——— (2000): "Semi-Parametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–177. [738]

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model," *Journal of Econometrics*, 3, 205–228. [738,745]

——— (1987): "Semi-Parametric Analysis of Random Effects Linear Models From Binary Panel Data," *Econometrica*, 55, 357–362. [738,741,742,744]

MCFADDEN, D. L. (1978): "Modeling the Choice of Residential Location," in *Spatial Interaction Theory and Residential Location*, ed. by A. Karlgvist et al. North-Holland. [740]

——— (1981): "Economic Models of Probabilistic Choice," in *Structural Analysis of Discrete Data With Econometric Applications*, ed. by C. Manski and D. McFadden. MIT Press. [740]

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Chapter 36 Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. L. McFadden. Elsevier. [759]

PAKES, A., AND J. PORTER (2013): "Moment Inequalities for Semi-Parametric Multinomial Choice With Fixed Effects," Working Paper, Harvard University. [738,741]

POWELL, J., AND P. RUUD (2008): "Simple Estimators for Semi-Parametric Multinomial Choice Models," Working Paper. [738]

ROCKAFELLAR, R. T. (1970): *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton, NJ: Princeton University Press. [739]

SHERMAN, R. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137. [751]

SHI, X., M. SHUM, AND W. SONG (2018): "Supplement to 'Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity'," *Econometrica Supplemental Material*, 86, https://doi.org/10.3982/ECTA14115. [744]

VILLANI, C. (2003): *Topics in Optimal Transportation*. American Mathematical Society, Graduate Studies in Mathematics, Vol. 58. [739]

Co-editor Elie Tamer handled this manuscript.

Manuscript received 21 January, 2016; final version accepted 28 September, 2017; available online 18 October, 2017.