DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

ESTIMATING MULTINOMIAL CHOICE MODELS USING CYCLIC MONOTONICITY

Xiaoxia Shi
University of Wisconsin-Madison

Matthew Shum
Caltech

Wei Song
University of Wisconsin-Madison

SOCIAL SCIENCE WORKING PAPER 1397

October 2014

Estimating Multinomial Choice Models using Cyclic Monotonicity

Xiaoxia Shi Matthew Shum Wei Song

Abstract

This paper proposes a new identification and estimation approach to semi-parametric multinomial choice models that easily applies to not only cross-sectional settings but also panel data settings with unobservable 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 restriction 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, and apply it to a panel data set to study the determinants of the demand of bathroom tissue.

JEL classification numbers: C14, C25

Key words: Cyclic Monotonicity, Multinomial Choice, Panel Data, Fixed Effects

Estimating Multinomial Choice Models using Cyclic Monotonicity*

Xiaoxia Shi† Matthew Shum‡ Wei Song§

1 Introduction

Consider a multinomial choice problem where an agent chooses from $K+1$ options (labelled $k = 0, \ldots, K$). Option k gives the agent indirect utility

$$\beta' X^k + \epsilon^k, \tag{1}$$

where X^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, and ϵ^k is an unobservable utility shock. The agent chooses the option that gives her the highest utility: $Y = k \Leftrightarrow k \in \text{argmax}_{k=0,\ldots,K} \left(\beta' X^k + \epsilon^k\right)$, where Y denotes the multinomial choice indicator.

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. We first show that the cyclic monotonicity property applies to the vector of choice probabilities $\left\{P(Y = k|X^0, \ldots, X^K)\right\}_{k=0,1,\ldots,K}$ emerging from any multinomial choice model, when viewed as a function of the vector of linear utility indices $(\beta' X^0, \ldots, \beta' X^K)'$. The

*Acknowledgment: We thank Bruce Hansen, Federico Echenique, Jack Porter, and seminar audiences at Northwestern, NYU, and UC Riverside for useful comments. 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.

†xshi@ssc.wisc.edu

‡mshum@caltech.edu

§wsong2@wisc.edu

cyclical monotonicity property implies a collection of of moment inequalities. These moment inequalities point identify the coefficient β under an appropriate normalization and straightforward assumptions on the covariates.

Importantly, our approach easily applies to (short) panel data models with fixed effects. For such models, we consider cyclic monotonicity with respect to cross-time period cycles, which leads to moment inequalities in which the fixed effects are differenced out. As reviewed below, this approach is one of the first nonlinear differencing techniques available for panel multinomial choice models. We then propose a consistent estimator for β, the computation of which requires only convex optimization.

Several other features of our approach are noteworthy. First, we do not require any assumption (beyond continuity) on the joint distribution of $(\epsilon^0, \ldots, \epsilon^K)'$. In particular, this joint distribution needs not be exchangeable with respect to the indices. Second, in the panel data context, we impose no restriction on the dependence between the covariates and the fixed effects. Particularly, we do not require the presence of a special regressor that is conditionally independent of the fixed effects. Finally, the moment inequalities that we derive apply to both the case of "aggregate data", in which only choice frequencies are observed across a large set of markets (this is a common data structure in the field of empirical industrial organization (IO)), as well as the case of "individual-level data", in which choices are observed for individual agents.

As a tradeoff, we assume the independence between the idiosyncratic error and the covariates. However, this independence can be relaxed when a control variable/function is available and also can be replaced by some other assumptions that guarantees the convexity of the social surplus function, as discussed in Section 2 below.

The paper proceeds as follows. In section 2, we introduce the notion of cyclic monotonicity and relate it to multinomial choice models. Subsequently, we present the moment inequalities emerging from cyclic monotonicity, and give assumptions under which these inequalities suffice to point identify the parameters of interest. Here we present numerical illustrations of how the identified set of β shrinks as the point identification assumptions get closer to being satisfied. Section 3 discusses the application to panel data multinomial choice models, with individual-specific fixed effects. Section 4 discusses estimation and contains a result about the consistency of our estimator.

The existing literature on multinomial choice models is voluminous, and Section 5 contains a full discussion that relates our approach to the literature. As an empirical

illustration, we apply our methods to estimate a market-level demand model for toilet tissue, using supermarket scanner data. This is discussed in section 6. Section 7 concludes.

2 Cyclic monotonicity and multinomial choice models

We begin by defining cyclic monotonicity, the central notion of this paper.

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

$$\sum_{j=1}^{J} (u_{j+1} - u_j)' f(u_j) \leq 0, \tag{2}$$

where $u_{J+1} = u_1$. 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.[1]

For real-valued functions defined on a real-space (i.e., $m = 1$), cyclic monotonicity is equivalent to monotonicity. In this sense, cyclic monotonicity is one way of generalizing monotonicity in a vector-valued context. We make use of the following basic result which relates cyclic monotonicity to convex functions (see e.g, Rockafellar (1970, Ch. 24), Villani (2003, Sct. 2.3)):

Proposition 1 (Cyclic monotonicity and Convexity)**.** *Consider a differentiable function $F : \mathcal{U} \to R$ for an open convex set $\mathcal{U} \subseteq R^m$. 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 $\mathcal{U}$.*

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. To convey the central ideas, we will stay in the cross-sectional setting throughout this section. The

[1] Technically, the definition 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."

results in the cross-sectional setting form the building blocks for the panel setting, which will be discussed in detail in the next section.

Define $U^k = \beta' X^k$, and $u^k = \beta' x^k$ for $k = 0, 1, \ldots, K$ for a generic realization $x^k \in \mathcal{X}$ of X^k. Also let $\vec{U} = (U^0, \ldots, U^K)'$, $\vec{u} = (u^0, u^1, \ldots, u^K)'$, $\vec{\epsilon} = (\epsilon^0, \ldots, \epsilon^K)'$, $\vec{x} = (x^{0,\prime}, \ldots, x^{K,\prime})'$ and $\vec{X} = (X^{0,\prime}, \ldots, X^{K,\prime})'$. We start with the social surplus function (or the expected utility of a representative agent making the multinomial choice decision):

$$\mathcal{G}(\vec{u}) = E\{\max_k [U^k + \epsilon^k] | \vec{U} = \vec{u}\}. \tag{3}$$

Now we introduce the assumption on the error distribution:
Assumption 2.1 (Error Distribution)**.** (a) $F_{\vec{\epsilon}|\vec{X}}(\cdot|\vec{X}) = F_{\vec{\epsilon}}(\cdot)$, *where* $F_{\vec{\epsilon}|\vec{X}}$ *is the conditional distribution of* $\vec{\epsilon}$ *given* $\vec{X}$*, and* $F_{\vec{\epsilon}}$ *is the marginal distribution of* $\vec{\epsilon}$*, and*

(b) $F_{\vec{\epsilon}}(\cdot)$ *is continuous everywhere.*

(c) *The support of* $\vec{U}$ *is a subset of that of* $\vec{\epsilon}$*.*

Remarks. Some words on part (a) of the assumption are worthwhile as it may appear to be a restrictive assumption. To begin, we stress that it is substantially weaker than the commonly used independent type-I extreme value (logit) assumption. The main advantage of part (a) over logit is that it allows arbitrary heterogeneity and dependence of ϵ^k across k, which is important to avoid the IIA (independence of irrelevant alternative) restriction on conditional choice probabilities.

Moreover, relaxing part (a) within our framework can be done in two ways. First, we can instead use a conditional version $F_{\vec{\epsilon}|\vec{X},Z}(\cdot|\vec{X}, Z) = F_{\vec{\epsilon}|Z}F(\cdot|Z)$ given a control variable/vector Z. Then all our subsequent identification and estimation analysis follows with the additional conditioning on Z; in particular, the cyclic monotonicity inequality restrictions (Eqs. (6) and (7) below) derived below will just need to hold with the additional conditioning event $Z = z$ for all possible values of z. Similar uses of control variables are common in the treatment effect literature (see e.g. Rosenbaum and Rubin (1983) and Imbens (2004)). Such control variables, if not available in the data, may be constructed as control functions after imposing a structure on the generation process of $\vec{X}$, similarly to the non-separable model literature (see e.g. Imbens and Newey (2009)). Second, part (a) may be weakened to other conditions that guarantee the convexity of $\mathcal{G}(\vec{u})$. For example, we may consider $\epsilon^k = g_k(U^k, \eta^k)$, where $(\eta^0, \ldots, \eta^K)'$ is independent of $\vec{X}$, and $g_k(\cdot, \eta^k)$ is convex for all k and all values of η^k. Fully exploring these extensions is beyond the scope of this paper, but is the topic of ongoing research. ■

Assumption 2.1 guarantees that $\mathcal{G}(\cdot)$ is convex and differentiable in R^{K+1}, and that the choice probability vector is the gradient of $\mathcal{G}$:

$$\vec{p}(\vec{u}) = \nabla \mathcal{G}(\vec{u}), \tag{4}$$

where $\vec{p}(\vec{u}) = E(\vec{Y}|\vec{U} = \vec{u})$ with $\vec{Y} = (Y^0, \ldots, Y^K)'$ and $Y^k = 1\{U^k + \epsilon^k \geq U^\ell + \epsilon^\ell, \ \forall \ell = 0, \ldots, K\}$. This is developed in the next Lemma, which shows the convexity and differentiability of $\mathcal{G}$ on R^{K+1}, as well as equation (4). The proof of the lemma is given in the appendix.[2]

Lemma 1 (Gradient)**.** *Suppose that Assumption 2.1 holds. Then*

(a) $\mathcal{G}(\cdot)$ *is convex on* R^{K+1},

(b) $\mathcal{G}(\cdot)$ *is differentiable on* R^{K+1}, *and*

(c) *equation* (4) *holds.*

An immediate corollary of Proposition 1 and Lemma 1 connects cyclic monotonicity to the multinomial choice probabilities. The proof of the corollary is omitted.

Corollary 1 (Cyclic monotonicity of choice probabilities)**.** *Suppose that Assumption 2.1 holds. Then* $\vec{p}(\vec{u})$ *is cyclic monotone on* R^{K+1}, *that is for any integer* $j \geq 2$ *and any length-*J *cycle* $\vec{u}_1, \ldots, \vec{u}_J, \vec{u}_1$, *we have*

$$\sum_{j=1}^{J} [(\vec{u}_{j+1} - \vec{u}_j)' \vec{p}(\vec{u}_j)] \leq 0 \tag{5}$$

The cyclic monotonicity of $\vec{p}(\vec{u})$ is the basis of our identification results, which we discuss in the next section.

2.1 Identification

Let $\vec{\mathcal{X}}$ denote the support of $\vec{X}$. The cyclic monotonicity of $\vec{p}(\vec{u})$ immediately implies the following identifying inequalities: for any integer $J \geq 2$, and any length-J cycle

[2]Part (c) of the lemma is the well known Williams-Daly-Zachary (WDZ) theorem. See McFadden (1978, 1981) for discussions and proofs. Our proof of Lemma 1 in the Appendix also includes a self-contained proof.

$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_J, \vec{x}_1$ in $\vec{\mathcal{X}}$, we have

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' x_{j+1}^k - \beta' x_j^k) E(Y^k | \vec{X} = \vec{x}_j) \right] \leq 0, \tag{6}$$

where $\vec{x}_{j+1} = \vec{x}_1$ and x_j^k is subvector of $\vec{x}_j$ that is composed of the $(d_x k + 1)$th to the $d_x(k+1)$th coordinate of $\vec{x}_j$.

Before proceeding, it is useful to simplify the inequalities using the fact that $Y^0 = 1 - \sum_{k=1}^{K} Y^k$. With this equation plugged in, (6) becomes

$$\sum_{j=1}^{J} \left[\sum_{k=1}^{K} (\beta'(x_{j+1}^k - x_{j+1}^0) - \beta'(x_j^k - x_j^0)) E(Y^k | \vec{X} = \vec{x}_j) \right] \leq 0, \tag{7}$$

This suggests that it is without loss of generality to normalize $x^0 = 0$, which we do for the rest of the paper. In most applications, product 0 is an imaginary "outside option" and $x^0 = 0$ is explicitly used.[3]

For each cycle, the inequalities restrict β to a halfspace of $\mathbb{R}^{d_x}$ (with boundary passing through the origin). The inequalities for all cycles thus restrict β to the intersection of all the halfspaces defined by those cycles. Therefore, our identifying inequalities clearly restrict β to a strict subset of $\mathbb{R}^{d_x}$; in other words, they at least partially identify β.

On the other hand, it is clear that the inequalities (6) (or (7)) provide no identifying information for the scale of β. If the first coordinate of X^k is unity for all k, the inequalities are also not informative about the first coordinate of β. Such nonidentification is innate to the multinomial choice model where the location and the scale of the error vector $\vec{\epsilon}$ are not normalized. In the literature, various ways of normalizations have been adopted. For our identification and estimation, we find it the most convenient to adopt the following normalization.

Assumption 2.2 (Normalization)**.** (a) $\max_{\ell=1,\ldots,d_x} |\beta_\ell| = 1$, *where* β_ℓ *is the* ℓ*th coordintate of* β*.*

(b) *For any* $k = 0, \ldots, K$, X^k *does not contain a unity coordinate.*

[3] In other applications, all products are inside products and the characteristics for each product are measured, in which case we can arbitrarily set one product be product 0 and let the x^k for product k be the characteristics of product k minus that of product 0.

Essentialy, part (a) of this assumption, which is the scale normalization, restricts attention to vectors β which lie on "the unit square".

2.1.1 Assumptions for Point Identification

Next, we analyze the further assumptions needed to guarantee point identification given the normalization. For convenience, we start with the binary choice case, which illustrates clearly the main argument for point identification. Afterwards we present the general result of interest, for multinomial choice models.

In the binary choice case, the inequalities (7) can be simplified to[4]

$$\sum_{j=1}^{J} \left[(\beta' x_{j+1}^1 - \beta' x_j^1) E(Y^1 | \beta' X^1 = \beta' x_j^1) \right] \leq 0. \tag{8}$$

That is, the conditional mean function $E(Y^1|\beta' X^1 = u)$ is cyclic monotone in u. Because this function maps from R to R, cyclic monotonicity is equivalent to monotonicity. Therefore, the set of identifying inequalities for the binary choice case is simply: for all $x_1^1, x_2^1 \in \mathcal{X}$ we have

$$(\beta' x_2^1 - \beta' x_1^1)[E(Y^1|X^1 = x_2^1) - E(Y^1|X^1 = x_1^1)] \geq 0. \tag{9}$$

Let $\mathcal{A} = \{(x_2^1 - x_1^1)[E(Y^1|X^1 = x_2^1) - E(Y^1|X^1 = x_1^1)] : \ x_2^1, \ x_1^1 \in \mathcal{X}\}$, and let $\mathcal{C}$ denote the convex cone generated by $\mathcal{A}$, that is,

$$\mathcal{C} = \left\{ \sum_{\ell=1}^{L} \lambda_\ell a_\ell : a_\ell \in \mathcal{A}, \lambda_\ell \in [0, \infty), \text{ for all } \ell = 1, \cdots, L; \ L = 1, 2, 3 \ldots \right\}. \tag{10}$$

While $\mathcal{A}$ needs not be convex, $\mathcal{C}$ is convex. Hence it can be characterized as the intersection of all the halfspaces in $\mathbb{R}^{d_x}$ which contain $\mathcal{C}$. In this regard, the critical assumption for point identification of β is that $\mathcal{C}$ itself is a halfspace.

Assumption 2.3 (Sufficient and Necessary)**.** *The closure of $\mathcal{C}$ is a halfspace in $\mathbb{R}^{d_x}$.*

Intuitively, if $\mathcal{C}$ is not a halfspace, then it is the intersection of at least two distinct halfspaces with boundary passing through the origin. Let two of these halfspaces be $\{a \in \mathbb{R}^{d_x} : b_1' a \geq 0\}$, and $\{a \in \mathbb{R}^{d_x} : b_2' a \geq 0\}$. Then because these halfspaces contain

[4]Recall that x^0 is normalized to zero.

$\mathcal{C}$, we must have $b_1'a \geq 0$ and $b_2'a \geq 0$ for all $a \in \mathcal{A}$, implying that both b_1, b_2 satisfy the inequalities in (9). Hence β is not point identified.

Let $\Delta\mathcal{X}$ denote the set $\{x_1 - x_2 : x_1, x_2 \in \mathcal{X}\}$. A sufficient condition for Assumption 2.3 is that $\Delta\mathcal{X}$ surrounds the origin from every direction, which is stated concisely in the assumption below.

Assumption 2.4 (Sufficient)**.** *The set* $\{\lambda a : \lambda \in \mathbb{R}, \lambda \geq 0, a \in \Delta\mathcal{X}\}$ *equals* $\mathbb{R}^{d_x}$.

Then we have the following result:

Theorem 1 (Point identification - Binary choice)**.** *Suppose that Assumption 2.1 holds.*
(a) *The parameter β is uniquely identified under the normalization Assumption 2.2 if and only if Assumption 2.3 holds.*

(b) *Assumption 2.3 holds if Assumption 2.4 holds.*

Remarks. (a) Assumption 2.3 is a rather weak assumption because it does not require any of the covariates to have full support or even to be continuous.

(b) The sufficient condition in Assumption 2.4 rules out discrete regressors (but allows mixed regressors). On the other hand, given that there is no discrete regressor, Assumption 2.4 is rather weak. One sufficient condition is that the support of each X^1 contains a d_x-dimensional hypercube of positive volume.

(c) In general, Assumption 2.4 does not rule out deterministic relationships between covariates. For example, it holds if $X^1 = (W, W^2)'$ for a random variable W and its square term W^2, as long as the support of W contains an interval (of positive length) on the real line. ■

Next we present our main point identification result, for the multinomial choice case. The main argument here is a straightforward generalization of the binary choice argument presented above.

For any integer $J \geq 2$, let

$$\mathcal{A}(J) = \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{j+1}^k - x_j^k) E(Y^k | \vec{X} = \vec{x}_j) : x_j^k \in \mathcal{X} \ \forall j = 1, \ldots, J, \ x_{J+1}^k = x_1^k \ \forall k = 1, \ldots, K \right\}. \tag{11}$$

Let $\overline{\mathcal{A}}(\bar{J}) = \cup_{J=2}^{\bar{J}} \mathcal{A}(J)$. And let $\overline{\mathcal{C}}(\bar{J})$ be the convex cone generated by $\overline{\mathcal{A}}(\bar{J})$. Then the following condition is necessary and sufficient for point identification based on the inequalities (7) for all cycles of length at most $\bar{J}$. Here $\bar{J}$ is allowed to be ∞.

Assumption 2.5 (Sufficient and Necessary)**.** *The closure of* $\overline{\mathcal{C}}(\bar{J})$ *is a half space in* $\mathbb{R}^{d_x}$*.*

The sufficient condition is based on length-2 cycles and uses the strategies of the binary choice case. For $k_* \in \{1, \ldots, K\}$, let $\mathcal{X}^{k_*}(\vec{x}^{-k_*})$ be the support set of X^{k_*} given that $\vec{X}^{-k_*} = \vec{x}^{-k_*}$, where $\vec{X}^{-k_*}$ is $\vec{X}$ with its $(d_x k_* + 1)$th to $(d_x k_* + d_x)$th elements removed. Let $\Delta\mathcal{X}^{k_*}(\vec{x}^{-k_*}) = \mathcal{X}^{k_*}(\vec{x}^{-k_*}) - \mathcal{X}^{k_*}(\vec{x}^{-k_*})$.

Assumption 2.6 (Sufficient)**.** *There exists* $k^* \in \{1, \ldots, K\}$ *and a vector* $\vec{x}^{-k_*} \in (\mathcal{X})^{d_x(K-1)}$ *such that the set* $\{\lambda a : \lambda \geq 0, a \in \Delta\mathcal{X}^{k_*}(\vec{x}^{-k_*})\}$ *equals* $\mathbb{R}^{d_x}$*.*

Then we have the following result.

Theorem 2 (Point Identification - Multinomial choice)**.** *Suppose that Assumption 2.1 holds.* (a) *The parameter* β *is uniquely identified by the identifying inequalities (7) for all cycles of length at most* $\bar{J}$ *under the normalization Assumption 2.2, if and only if Assumption 2.5 holds.*

(b) *Assumption 2.5 holds with* $\bar{J} = 2$ *if Assumption 2.6 holds.*

The proof of part (a) of the theorem is the same as that of Theorem 1(a) and the proof of part (b) of the theorem is the same as that of Theorem 1(b) once we condition on the event $\vec{X}^{-k_*} = \vec{x}^{-k_*}$. Thus, the proof of this theorem is omitted.

2.2 Numerical Illustration

In this subsection, we consider two numerical examples to illustrate the identifying power of cyclical monotonicity. We consider a binary choice model followed by a three-choice multinomial model.

2.2.1 Binary Choice Model

First, we consider a numerical example for the binary choice model, where the regressors X^1, are a 3-dimensional vector with support $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$. The covariate value for option 0, X^0, is normalized to zero.

We generate the data by setting the error term $\{\epsilon^k\}_{k=0}^K$ to be independently type 1 extreme-value distributed across both choices, leading to binary logit choice probabilities. We also set the true value $\beta = (1, 1, 1)'$. Using this information, we can precisely compute $E(Y^1|X^1 = x^1)$ for given value x^1 of X^1.

By the discussion in Section 2.1.1, if we allow X^1 to be a continuous random vector whose support contains a hypercube, the identifying inequalities in (9) point identify β. But if X^1 only takes a finite number of values, the inequalities (9) may only restrict β to an identified set. Below, we vary the support of X^1 to demonstrate how the identification power of (9) increases as we increase the number of points in the support of X^1.

For comparison purpose, we also consider the identification power of the median-independence condition underlying the maximum score estimator (Manski, 1975). The identifying restriction of median-independence is described in Section 5 below.

We consider four designs of $\mathcal{X}$:

[A.] $\mathcal{X} = \{0, 0.5, 1\}^3$,

[B.] $\mathcal{X} = \{0, 0.25, 0.5, 0.75, 1\}^3$,

[C.] $\mathcal{X} = \{0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\}^3$,

[D.] $\mathcal{X} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}^3$.

We normalize $\beta_1 = 1$, and plot the identified set for (β_2, β_3). Figure 1 below shows the identified set defined by the inequalities (9) for all four designs, as well as the identified set defined by the maximum score identification condition (30). As we can see, the identified set shrinks rapidly as we increase the density of points in $\mathcal{X}$, and at the most dense design (Design D), the identified set is numerically indistinguishable from the singleton point $(1, 1)$.

In contrast, the identified set for the maximum score approach, for all four designs, remains the entire first quadrant $\{\beta_1 \geq 0, \beta_2 \geq 0\}$. This is not surprising given that the maximum score condition is derived under a weaker median-independence assumption. When its weaker assumption on the error distribution is not accompanied by the full support of some of the regressors, evidently, identification becomes quite weak, as in this example.

2.2.2 Multinomial (three choice) Model

Now we consider a multinomial (three choice) model. Each individual's choice set consists of three elements $k \in \{0, 1, 2\}$ where choices 1 and 2 are characterized by covariates

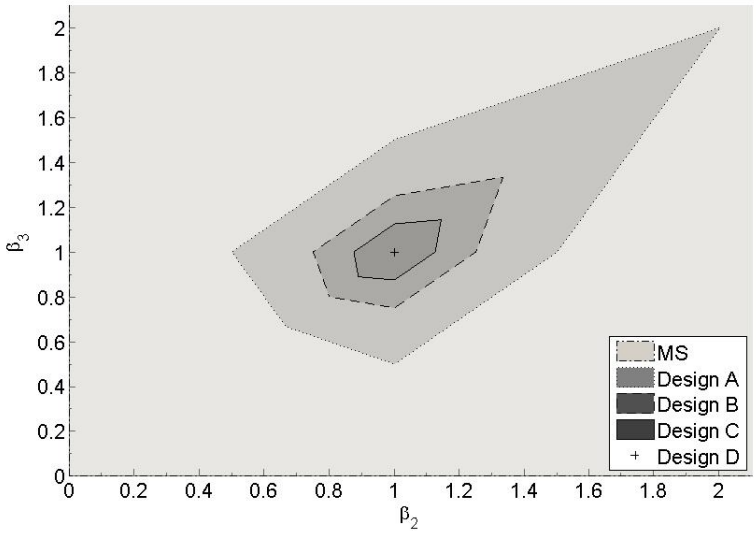

Figure 1: **Illustration of Identified Sets in Binary Choice Example.** Designs A-D show the identified set of the cyclic monotonicity inequalities (9) under the four designs specified in the text. MS shows the identified set of the maximum score identification condition (30), which is first quadrant for all four designs.

$X^1 = (X_{11}, X_{12}, X_{13})'$ and $X^2 = (X_{21}, X_{22}, X_{23})'$, respectively. The covariate for X^0 for option 0 is normalized to zero.

As in the binary choice model, we consider a multinomial logit specification for the utility shocks ϵ_i^k. We consider the same four designs as in the binary example, and normalize β_1 to 1. Figure 2 shows the identified set of the inequalities (7) with length-2 cycles for the four designs. As we can see, the identified set shrinks as the points become dense in the support of X^k, $k = 1, 2$. At the most dense design, the identified set is numerically indistinguishable from a singleton point at the true value of $\beta_2 = \beta_3 = 1$.[5]

[5]Notice that the shape of the identified sets in Figure 2 is not as nice as in Figure 1. This is because the identified sets for the binary choice model are plotted precisely using all possible cycles. On the other hand, the number of "all possible cycles" is prohibitively large in the trinary example except in the most sparse design. Thus, the identified sets plotted in Figure 2 are approximations of the identified set using a sample of 5000 cycles. To draw the cycle sample, we assume that $(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23})$ are i.i.d., and each assigns equal probability to its support points. To form a random length-2 cycle, we draw the two points from the distribution just specified independently, and form the cycle using these two points.

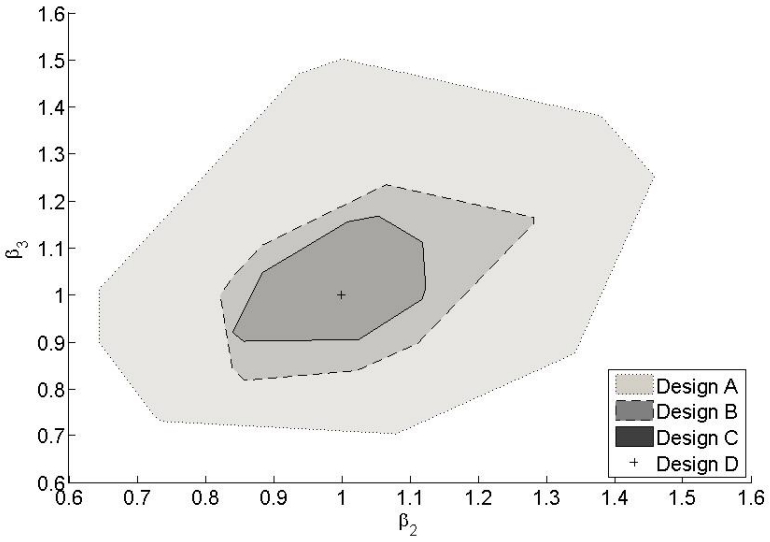

Figure 2: **Illustration of Identified Sets in Trinary Choice Example.** Designs A-D show the identified set of the cyclic monotonicity inequalities (9) under the four designs specified in the text.

3 Panel Data Multinomial Choice Models

One notable advantage of the cyclic monotonicity framework is the immediate applicability to panel multinomial choice models with fixed effects. In this section, we start with the aggregated panel data setting (explained below) because that is the data structure in our empirical applications. We also discuss the standard individual panel data setting in order to connect to the broader semi-parametric panel discrete choice literature.

3.1 Aggregated Panel Data

The aggregate panel data structure is often encountered in empirical IO. Typically, 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 $(\vec{S}_{ct}, \vec{X}_{ct})_{c=1\,t=1}^{C\quad T}$ where $\vec{X}_{ct} = (X_{ct}^{0,\prime}, \ldots, X_{ct}^{K,\prime})'$ is the region/time-level covariate, and $\vec{S}_{ct} = (S_{ct}^0, \ldots, S_{ct}^K)'$, S_{ct}^k is the fraction of n_{ct} agents in region c and time t who chose option k, i.e.

$$\vec{S}_{ct} = n_{ct}^{-1} \sum_{i=1}^{n_{ct}} \vec{Y}_{ict}, \tag{12}$$

where $\vec{Y}_{ict} = (Y_{ict}^0, \ldots, Y_{ict}^K)'$ denotes the vector of choice indicators for individual i in region c and time t. Only a "short" panel is required, as our approach works with as few as two periods. Suppose that the support of X_{ct}^k for all k, c, t is $\mathcal{X} \subseteq R^{d_x}$, and the support of $\vec{X}_{ct}$ is $\vec{\mathcal{X}}$.

We model the individual choice $\vec{Y}_{ict}$ as

$$Y_{ict}^k = 1\{\beta' X_{ct}^k + \alpha_c^k + \epsilon_{ict}^k \geq \beta' X_{ct}^{k'} + \alpha_c^{k'} + \epsilon_{ict}^{k'} \ \forall k'\}. \tag{13}$$

where $\vec{a}_c = (a_c^0, \ldots, a_c^K)'$ is the choice-specific regional fixed effect, and $\vec{\epsilon}_{ict} = (\epsilon_{ict}^0, \ldots, \epsilon_{ict}^K)'$ is the vector of idiosyncratic shocks. We make the following assumptions

Assumption 3.1 (Stationarity)**.** *The vector of utility shocks $\vec{\epsilon}_{ict}$ is identically distributed across t.*

Assumption 3.2 (Contemporaneous Exogeneity)**.** (a) *The vector of utility shocks $\vec{\epsilon}_{ict}$ is exogenous in the sense of: $F_{\vec{\epsilon}_{ict}|\vec{X}_{ct},\vec{a}_c} = F_{\vec{\epsilon}_{ict}}$.*

(b) *$F_{\vec{\epsilon}_{ict}}(\cdot)$ is continuous everywhere.*

(c) *The support of $\beta' X_{ct}^k + a_c^k$ is a subset of that of ϵ_{ict}^k for all $k = 0, \ldots, K$.*

Consider cycles on $\vec{\mathcal{X}}$, the support of $\vec{X}_{ct} := (X_{ct}^{0,\prime}, \ldots, X_{ct}^{K,\prime})'$. We cannot use cycles on the support of $(X_{ct}^{0,\prime}, a_c^0, \ldots, X_{ct}^{K,\prime}, a_c^K)'$ because we want to derive inequalities where the fixed effects $\vec{a}_c := (a_c^0, \ldots, a_c^K)'$ are held constant, so that they can be "differenced" out. Using Assumption 3.2, for all cycles $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_J, \vec{x}_1 \in \vec{\mathcal{X}}$ and for all $J \geq 2$, we obtain, for any i, c, t:

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' x_{j+1}^k - \beta' x_j^k) E(Y_{ict}^k | \vec{X}_{ct} = \vec{x}_j, \vec{a}_c) \right] \leq 0, \tag{14}$$

where $\vec{x}_{i(J+1)} = \vec{x}_{i1}$ and x_j^k is the subvector of $\vec{x}_j$ that is composed of the $(d_x k + 1)$th to the $d_x(k+1)$th coordinate of $\vec{x}_j$.

Further applying Assumption 3.1, we can write (14) equivalently as: for all $J \geq 2$, any length-J cycles of time indices $t_1, \ldots, t_J, t_{J+1} \equiv t_1 \in \{1, \ldots, T\}$ and any c,

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' X_{t_{j+1}}^k - \beta' X_{t_j}^k) E(Y_{ict_j}^k | \vec{X}_{ct_j}, \vec{a}_c) \right] \leq 0, \ a.s. \tag{15}$$

In aggregated panel data sets, we can expect $\vec{S}_{ct}$ to be a good estimator of $E(Y_{ict}^k | \vec{X}_{ct}, \vec{a}_c)$

uniformly over c and t. In this case, inequalites (15) (and equivalently (14)) serve directly as our identifying inequalities even though $\vec{a}_c$ is unobserved.

The identifying inequalities (14) are very similar to the general case (6) except that the choice probability is now conditional on both $\vec{X}_{ct}$ and $\vec{a}_c$. Since these conditional choice probabilities can be well-estimated directly from the data in the aggregate setting, this difference does not affect any of the identification results given above. Thus, Theorems 1 and 2 still apply with appropriate notational adjustment.

3.2 Individual Panel Data

In an individual panel data set, we observe $(\vec{Y}_{it}, \vec{X}_{it})_{i=1}^{n}{}_{t=1}^{T}$ for individual i and time t, where $\vec{Y}_{it} = (Y_{it}^0, \ldots, \vec{Y}_{it}^K)'$, and $\vec{X}_{it} = X_{it}^{0,'}, \ldots, X_{it}^{K,'})'$. We model the individual choice $\vec{Y}_{it}$ as

$$Y_{it}^k = 1\{\beta' X_{it}^k + a_i^k + \epsilon_{it}^k \geq \beta' X_{it}^{k'} + a_i^{k'} + \epsilon_{it}^{k'};\ \forall k'\} \tag{16}$$

where ϵ_{it}^k, $k = 0, \ldots, K$ are the idiosyncratic shocks. In the individual-level dataset, the fixed effects a_i^k are choice- and individual-specific effects. This is an important difference vis-a-vis the agrgegate model in the previous section, where the a_c^k are choice- and *region*-specific fixed effects.

Analogously to the aggregated panel data setting above, we obtain the following inequalities: for any integer $J \geq 2$, and any length-J cycle $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_J, \vec{x}_1 \in \vec{\mathcal{X}}$, we have, for any i, t,

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' x_{j+1}^k - \beta' x_j^k) E(Y_{it}^k | \vec{X}_{it} = \vec{x}_j, \vec{a}_i) \right] \leq 0, \tag{17}$$

where $\vec{x}_{i(J+1)} = \vec{x}_{i1}$ and x_j^k is the subvector of $\vec{x}_j$ that is composed of the $(d_x k + 1)$th to the $d_x(k+1)$th coordinate of $\vec{x}_j$.

The difficulty with individual-level data is that the data do not provide a consistent estimator for the individual-specific conditional choice probabilities $E(Y_{it}^k | \vec{X}_{it} = \vec{x}_j, \vec{a}_i)$. This is a critical difference vis-a-vis the aggregate-level data case; here, since we only observe several choices per individual, we cannot "difference out" the individual-specific fixed effect. Hence we proceed by making stronger stochastic assumptions which enable us to "integrate out" the fixed effects:

Assumption 3.3 (Stationarity)**.** *The vector of utility shocks $\vec{\epsilon}_{it}$ is identically distributed across t.*

Assumption 3.4 (Strict exogeneity)**.** (a) *The vector of utility shocks $\vec{\epsilon}_{it}$ is strictly exogenous in the sense of: $(\vec{\epsilon}_{it})_{t=1}^{T} \perp (\vec{X}_{it})_{t=1}^{T}, \vec{a}_i$.*

(b) *$F_{\vec{\epsilon}_{it}}(\cdot)$ is continuous everywhere.*

(c) *The support of $\beta' X_{it}^k + a_i^k$ is a subset of that of ϵ^k for all $k = 0, \ldots, K$.*

Remarks. (a) The stationarity assumption and the strict exogeneity assumption together imply Manski's (1987) and Pakes and Porter's (2013) conditional homogeneity assumption, and in fact is the leading sufficient condition mentioned in the latter. The strict exogeneity condition is stronger than the contemporaneous exogeneity required for in the aggregated panel setting, and is needed here for integrating out the fixed effect in the identifying conditions.

(b) We allow the endogeneity of the covariates in the form of correlation between X and the unobserved fixed effect because the dependence between $\vec{a}_i$ and $\{\vec{X}_{it}\}_t$ is unrestricted. ■

To see how the stronger assumptions help, note that Assumption 3.3 implies, for all t,

$$E(\vec{Y}_{it}|\vec{X}_{it} = \vec{x}, \vec{a}_i) = E(\vec{Y}_{i1}|\vec{X}_{i1} = \vec{x}, \vec{a}_i). \tag{18}$$

Thus, the inequalities (17) can be written as

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' x_{j+1}^k - \beta' x_j^k) E(Y_{it_j}^k | \vec{X}_{it_j} = \vec{x}_j, \vec{a}_i) \right] \leq 0, \tag{19}$$

where $\{t_1, \ldots, t_J\}$ is the set of distinct time indices. Also, Assumption 3.4 implies that $E(Y_{it_j}^k | \vec{X}_{it_j} = \vec{x}_j, \vec{a}_i) = E(Y_{it_j} | \vec{X}_{it_1}, \ldots, \vec{X}_{it_{j-1}}, \vec{X}_{it_j} = \vec{x}_j, \vec{X}_{it_{j+1}}, \ldots, \vec{X}_{it_J}, \vec{a}_i)$. Therefore, (19) can be written as

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' x_{j+1}^k - \beta' x_j^k) E(Y_{it_j}^k | \vec{X}_{it_1} = \vec{x}_1, \ldots, \vec{X}_{it_J} = \vec{x}_J, \vec{a}_i) \right] \leq 0. \tag{20}$$

Now taking conditional expectations over the distribution of $\vec{a}_i$ given $\vec{X}_{it_1}, \ldots, \vec{X}_{it_J}$ on both sides of the above inequalities, we obtain for all $J \geq 2$ and any length-J cycle of

time indices $t_1, \ldots, t_J, t_{J+1} \equiv t_1$,

$$\sum_{j=1}^{J} \left[\sum_{k=0}^{K} (\beta' X_{t_{j+1}}^k - \beta' X_{t_j}^k) E(Y_{it_j}^k | \vec{X}_{it_1}, \ldots, \vec{X}_{it_J}) \right] \leq 0, \ a.s. \tag{21}$$

Therefore, in the panel data setting with individual data, our identifying conditions are inequalities in (21).

4 Estimation and Inference

4.1 Cross-sectional Case

In this subsection, we assume that β is point identified based on the inequalities Eq. (6) constructed using cycles of length at most $\bar{J}$.

We have a data set $\{(\vec{X}_i, \vec{Y}_i)\}_{i=1}^n$. Based on this data set, suppose that there is a uniformly consistent estimator $\hat{p}^k(\vec{x})$ for $E(Y^k | \vec{X} = \vec{x})$ for all k; for example, this can be a kernel regression estimator. Then a consistent estimator of β can be obtained as

$$\widehat{\beta} = \arg \min_{\substack{b=(b_1,\ldots,b_{d_x})': \\ \max_{\ell=1,\ldots,d_x} |b_\ell|=1}} Q_n(b), \tag{22}$$

where

$$Q_n(b) = \max_{J=2,\ldots,\bar{J}} \max_{\substack{1 \leq i_1,\ldots,i_J \leq n \\ i_{J+1}=i_1}} \left[\frac{J^{-1} \sum_{j=1}^{J} \sum_{k=0}^{K} (b' X_{i_{j+1}}^k - b' X_{i_j}^k) \hat{p}^k(\vec{X}_{i_j})}{\max_{j=1,\ldots,J, k=0,\ldots,K} \|X_{i_j}^k\|} \right]_+, \tag{23}$$

where $[a]_+ = \max\{0, a\}$. The weight $\max_{j=1,\ldots,J,k=0,\ldots,K} \|X_{i_j}^k\|$ is used to normalize the scale of X_j^k's in each cycle. It allows the consistency result to be shown without assuming that $\mathcal{X}$ is a bounded set. If $\mathcal{X}$ indeed is bounded, then removing the weight does not affect consistency.

Let the population version of $Q_n(b)$ be

$$Q(b) = \max_{J=2,\ldots,\bar{J}} \sup_{\vec{x}_1,\ldots,\vec{x}_J \in \vec{\mathcal{X}}} \left[\frac{J^{-1} \sum_{j=1}^{J} \sum_{k=0}^{K} (b' x_{j+1}^k - b' x_j^k) E(Y^k | \vec{X} = \vec{x}_j)}{\max_{j=1,\ldots,J, k=0,\ldots,K} \|x_j^k\|} \right]_+ \tag{24}$$

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

Theorem 3 (Consistency)**.** *Suppose that the following conditions hold:*

(i) $Q(b)$ *is uniquely minimized at* $b = \beta$.

(ii) $\sup_{\vec{x}\in\vec{\mathcal{X}}} \sum_{k=1}^{K} |\hat{p}^k(\vec{x}) - E(Y^k|\vec{X} = \vec{x})| \to_p 0$ *as* $n \to \infty$.

(iii) $p^k(\vec{x}) \equiv E(Y^k|\vec{X} = \vec{x})$ *is continuous in* $\vec{x}$ *on* $\vec{\mathcal{X}}$.

Then, we have $\widehat{\beta} \to_p \beta$.

While we have derived consistency for the estimator, calculations for the limit distributions are difficult, due to the "one-sided" nature of the objective function. We leave it for future work.

4.2 Individual Panel Data

In this subsection, we consider a panel data set $\{\vec{X}_{it}, \vec{Y}_{it})\}_{i=1}^{n}{}_{t=1}^{T}$, and let β be point identified based on the inequalities Eq. (21) constructed using cycles of length up to $\bar{J}$.

Assume that we have a short panel, that is, in the asymptotic analysis, T is fixed and $n \to \infty$. Based on the panel data set, suppose that there is a uniformly consistent estimator $\hat{p}^k_{t_j|t_1,\ldots,t_J}(\vec{x}_1, \ldots, \vec{x}_J)$ for $E(Y^k_{it_j}|\vec{X}_{it_1} = \vec{x}_1, \ldots, \vec{X}_{it_J} = \vec{x}_J)$ for all k, all $j = 1, \ldots, J$, and all $t_1, \ldots, t_J$. For example, this can be a kernel regression estimator. Then a consistent estimator of β can be obtained as

$$\widehat{\beta} = \arg \min_{b=(b_1,\ldots,b_{d_x})':\max_\ell |b_\ell|=1} Q_n(b), \qquad \text{where } Q_n(b) = \tag{25}$$

$$\max_{J=2,\ldots,\bar{J}} \max_{\substack{1\le t_1,\ldots,t_J\le T,\\ t_{J+1}=t_1}} \max_{1\le i\le n} \left[\frac{J^{-1}\sum_{j=1}^{J}\sum_{k=0}^{K}(b'X^k_{it_{j+1}} - b'X^k_{it_j})\hat{p}^k_{t_j|t_1,\ldots,t_J}(\vec{X}_{it_1}, \ldots, \vec{X}_{it_J})}{\max_{j=1,\ldots,J;k=0,\ldots,K} \|X^k_{it_j}\|} \right]_+ \tag{26}$$

The estimator is consistent by similar arguments as those for Theorem 3 given that the following conditions hold: (i) the inequalities (21) with $J = 2, \ldots, \bar{J}$ together point identify β; (ii) $\sup_{\vec{x}_{j'}\in\vec{\mathcal{X}}\ \forall j'=1,\ldots,J} \sum_{k=1}^{K} |\hat{p}^k_{t_j|t_1,\ldots,t_J}(\vec{x}_1, \ldots, \vec{x}_J) - E(Y^k_{it_j}|\vec{X}_{it_1} = \vec{x}_1, \ldots, \vec{X}_{it_J})| \to_p 0$ as $n \to \infty$ for all $j = 1, \ldots, J$, all $t_1, \ldots, t_J \in \{1, \ldots, T\}$, and all $J = 2, \ldots, \bar{J}$; and (iii) $E(Y^k_{it_j}|\vec{X}_{it_1} = \cdot, \ldots, \vec{X}_{it_J} = \cdot)$ is continuous for all $k = 0, \ldots, K$, all $j = 1, \ldots, J$, all $t_1, \ldots, t_J \in \{1, \ldots, T\}$, and all $J = 2, \ldots, \bar{J}$.

For the same reason as the previous subsection, we leave the limiting distribution of $\widehat{\beta}$ for future research.

4.3 Aggregated Panel Data

When an aggregated panel data set like that discussed in Section 3.1 is available, estimating the conditional choice probability is easier. This is because, we can use $\vec{S}_{ct}$ to estimate $E(\vec{Y}_{ict}|\vec{X}_{ct}, \vec{a}_c)$. If $\inf_{c,t} n_{ct}$ grows fast enough with $C \times T$, this estimator is uniformly consistent, i.e.

$$\sup_c \sup_t \sup_J \sup_{t_1,\ldots,t_J \leq T} \|\vec{S}_{ct} - E(\vec{Y}_{ict}|\vec{X}_{ct}, \vec{a}_c)\| \to_p 0. \tag{27}$$

Section 3.2 of Freyberger's (2013) arguments (using Bernstein's Inequality) imply that the above convergence holds if $\log(C \times T)/\min_{c,t} n_{ct} \to 0$.

Given this, we can define the consistent estimator of β as

$$\widehat{\beta} = \arg \min_{b:\max_\ell |b_\ell|=1} Q_n(b), \tag{28}$$

where

$$Q_n(b) = \max_{J=2,\ldots,\bar{J}} \max_{c=1,\ldots,C} \max_{\substack{1 \leq t_1,\ldots,t_J \leq T, \\ t_{J+1}=t_1}} \left[\frac{J^{-1} \sum_{j=1}^{J} \sum_{k=0}^{K} (b' X^k_{ct_{j+1}} - b' X^k_{ct_j}) S^k_{ct_j}}{\max_{j=1,\ldots,J,k=0,\ldots,K} \|X^k_{ct_j}\|} \right]_+ . \tag{29}$$

This estimator is consistent by similar arguments as those for Theorem 3 if (i) the identifying inequalities (15) for $J = 2, \ldots, \bar{J}$ point identify β, (ii) $\log(C \times T)/\min_{c,t} n_{ct} \to 0$, and (iii) $E(\vec{Y}_{ict}|\vec{X}_{ct} = x, \vec{a}_c)$ is continuous in x almost surely for all c, t.

Note that the criterion function $Q_n(b)$ is a convex function of b in all three data environments. Obtaining a global minima of the convex function over a convex set typically is easy. This is why we use the normalization $\max_\ell |b_\ell| = 1$ instead of the perhaps more familiar normalization $\|b\| = 1$ in Assumption 2.2(a). Even though $\{b : \max_\ell |b_\ell| = 1\}$ is not a convex set, it can be written as the union of $2d_x$ convex sets: $\{b : b_1 = 1\}, \{b : b_1 = -1\}, \ldots, \{b : b_{d_x} = 1\}, \{b : b_{d_x} = -1\}$. Thus, obtaining $\widehat{\beta}$ amounts to solving $2d_x$ convex problems, which is more convenient in practice than solving a non convex problem with unknown numbers of local minima.

Throughout this section, we have focused on the estimation under the assumption that the conditions specified above for point identification are satisfied. In the case that these conditions are not satisfied, the parameters will only be point identfied, and we can consider an alternative inferential approach for this case based on recent work by Freyberger and Horowitz (2013). Since this approach is quite different in spirit to the methods described so far, we do not discuss it here.[6]

5 Comparison with Other Approaches

In this section, we compare our approach to other approaches to semi-parametric discrete choice models. Our approach relies on an assumption of independence between $(\epsilon^0, \ldots, \epsilon^K)$ and $(X^0, \ldots, X^K)$. For the special case of binary choice, our independence assumption is stronger than the median-independence assumption underlying the maximum score approach in Manski (1975, 1988). In maximum score, it is assumed that $med(\epsilon|X) = med(\epsilon) = 0$, which implies the maximum score identification condition:

$$E(Y^1|X^1 = x^1) \geq 0.5 \Leftrightarrow \beta' x^1 \geq 0, \quad \forall x^1 \in \mathcal{X}. \tag{30}$$

Suppose that the normalization $med(\epsilon) = 0$ is also used (instead of Assumption 2.2(b)). Then (30) is the cyclic monotnonicity applied to the cycles of the form $x^1, 0, x^1$. Our identifying inequalities use all cycles, and thus provides more restriction on the parameter.[7]

Still in the binary choice case, the cyclical monotonicity conditions are equivalent to the identification conditions implied by Han's (1987) monotone single index assumption. In fact, the identification inequalities (9) can be derived from Han's assumptions instead of the full independence assumptions that we are making. The former allows certain forms of heteroskedasticity. Nonetheless, we establish different point identification conditions than Han (1987) and propose a different estimator. Moreover, Han's approach only applies to binary choice; indeed, since cyclic monotonicity is one generalization of monotonicity to a multivariate setting, our approach may be considered one way of generalizing Han's estimator to the multinomial case.

[6] Please contact the authors for more details and empirical illustrations of this alternative approach.

[7] To supplement the median-independence assumption, the maximum score approach imposes some of the regressors to have full support in order to achieve point identification. As the numerical example in Section 2.2.1 illustrated, our approach has more identifying power than the maximum score approach when these support conditions are not satisfied.

The maximum score approach can be applied to multinomial choice models, under an additional symmetry (exchangeability) assumption on the joint distribution of $\vec{\epsilon}$ (e.g. Manski (1975), Fox (2007), Yan (2013)). Our independence assumption is neither stronger nor weaker than the symmetry assumption. Furthermore, our approach does not rely on large support for any of the covariates, does not impose a rank-order property between choice probabilities and utility indices, nor require that the sign of an coordinate of β is known, unlike some existing strategies for multinomial choice models (e.g., Lewbel (2000), Fox (2007)).

Klein and Spady (1993) provide a maximum likelihood based approach for semi-parametric binary choice models. Ichimura and Lee (1991), Lee (1995), as well as Ai's (1997) general semi-parametric likelihood approach may be applied to the estimation of a multinomial choice model like ours. However, it does not seem easy to extend the maximum likelihood approach to a panel data setting. Moreover, our approach involves solving convex minimization problems and thus has some computational advantage.

The literature on panel discrete choice models is smaller. Manski (1987), Honoré and Kyriazidou (2000) and Honoré and Lewbel (2002) propose alternative approaches for the *binary* choice model. These approaches do not immediately apply to multinomial models. Moreover, they impose various shape support assumptions on the covariates. On the other hand, we impose no restriction on the dependence between the covariates and the fixed effects, and do not require the presence of a special regressor that is conditionally independent of the fixed effects.

For panel data multinomial choice models, the on-going work of Pakes and Porter (2013) proposes a different nonlinear differencing approach. Since this is the only other approach available for the panel data setting with individual fixed effects for multinomial models, we describe it briefly here for comparison purpose. Pakes and Porter (2013) take a pair of time periods (t, s), and consider the difference

$$\delta_{i(t,s)}(\beta) := \vec{U}_{it} - \vec{U}_{is} \equiv (0, \beta'(X_{it}^1 - X_{is}^1), \cdots, \beta'(X_{it}^K - X_{is}^K))'. \tag{31}$$

Then they rearrange the $K+1$ coordinates of $\delta_{i(t,s)}(\beta)$ in descending order. Suppose that $k(0,\beta), \cdots, k(K,\beta)$ are respectively the index of the largest, second largest, ..., $(K+1)th$ largest coordinates of $\delta_{i(t,s)}(\beta)$. They show that for all $\ell \in \{0, \cdots, K\}$, the

following inequality holds

$$E\left[\sum_{v=0}^{\ell}(Y_{it}^{k(v,\beta)} - Y_{is}^{k(v,\beta)})\middle| \vec{X}_{it}, \vec{X}_{is}\right] \geq 0. \tag{32}$$

The difference between Eq. (21), which underlies our moment inequalities, and Eq. (32), which encapsulates the Pakes and Porter approach, is apparent. Neither set of moment inequalities nests the other. One obvious feature of our moment inequalities is that the linearity in β is preserved, which may simplify the estimation and inference.

6 Empirical Illustration

Here we consider an empirical illustration. We estimate a discrete choice demand model for toilet tissue, using store/week-level scanner data from different branches of Dominicks supermarket.[8] The toilet tissue category is convenient because there are relatively few brands of toilet paper, which simplifies the analysis. The data are collected at the store and week level, and report sales and prices of different brands of toilet tissue. For each of 54 Dominicks stores, we aggregate the store-level sales of toilet tissue up to the largest six brands, lumping the remaining brands into the seventh good (see Table 1).

Table 1: 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 estimate a panel version of the multinomial choice model, forming moment conditions based on cycles over weeks, for each store. In the estimation results below, we consider cycles of length 2.[9] Since data are observed at the weekly level, we consider subsamples of 5 weeks, 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

[8] 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.

[9] We find that using cycles of length up to 3 or 4 gives exactly the same estimates.

that can be formed from those weeks are used. In the estimation, the normalization in Assumption 2.2 is used; thus, no sign assumption is used for any coefficient. All signs are estimated from the data.

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 toilet tissue, while DEAL is defined as whether a given brand was on sale in a given store-week.[10] 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 2. One robust observation from the table is that the sign of the interaction term is positive, indicating that consumers are less sensitive to price when a product is red-tagged. This may be consistent with a "bounded-rationality" view of consumer behavior, whereby consumers may be less aware of a product's exact price once they are aware that it is on sale.

Table 2: Point Estimates for Demand Application

		5 week data	10 week data	15 week data
β_1	deal	.0267	.0294	-.0707
β_2	price	-1	-1	-1
β_3	price*deal	.0431	.0011	.1969

To compare the relative magnitude of the effects of PRICE and DEAL, consider the results using the 5-week data, and consider a product that has the highest price (about \$0.6, see Table 3) and is not on sale. The results imply that toggling DEAL from zero to one has the same effect on utility as a price drop of \$0.0533 ($= 0.0267 + 0.0431 * 0.6162$), which is 0.626 of the standard deviation in price. The corresponding numbers are \$0.0301 (or 0.3434 standard deviation), and \$0.0514 (or 0.5792 standard deviation) using the 10-week and the 15-week data results, respectively. These are small numbers, and suggest

[10]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 Assumption 2.6, which ensures point identification, plausible for this example.

that deal status is relatively unimportant, as compared to price, in its effect on consumer behavior.

Table 3: Summary Statistics

		min	max	mean	median	std.dev
5 week data	DEAL	0	1	0.4394	0	0.4831
	PRICE	0.1776	0.6162	0.3686	0.3625	0.0851
10 week data	DEAL	0	1	0.4350	0	0.4749
	PRICE	0.1776	0.6200	0.3637	0.3541	0.0876
15 week data	DEAL	0	1	0.4488	0	0.4845
	PRICE	0.1849	0.6200	0.3650	0.3532	0.0887

7 Conclusions

In this paper we explored how the notion of cyclic monotonicity can be exploited for the identification and estimation of multinomial choice models. 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. Moreover, in this paper we have mainly focused on estimation for the case when the researcher has aggregate panel data on choice probabilities; we also plan to explore estimation when only individual-level panel or cross-sectional data are available.

References

[1] C. Ai. A Semiparametric Maximum Likelihood Estimator. *Econometrica*, 65: 933-963.

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

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

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

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

[6] H. Ichimura and L. F. Lee. Semiprametric Estimation of Multiple Index Models: Single Equation Estimation. In *Nonparametric and semiparametirc methods in econometrics and statistics*, ed. W. Barnett, J. Powell, and G. Tauchen. Cambridge University Press, 1991.

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

[8] G. W. Imbens and W. K. Newey. Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77(5):1481-1512, 2009.

[9] R. Klein and R. Spady. An Efficient Semiparametri Estimator of Binary Response Models. *Econometrica*, 61:387-421, 1993.

[10] L. F. Lee. Semiparametric maximum likelihood estimation of polychotomus and sequential choice models. *Journal of Econometrics*, 65:381-428, 1995.

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

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

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

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

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

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

[17] A. Pakes and J. Porter. Moment Inequalities for Semiparametric Multinomial Choice with Fixed Effects. Working paper, Harvard University, 2013.

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

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

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

[21] J. Yan. Semiparametric Estimation of Multinomial Discrete Choice Models with Endogenous Regressors. Working paper, Chinese University of Hong Kong, 2013.

Proof. [Proof of Lemma 1]

(a) By Assumption 2.1(a), we have

$$\mathcal{G}(\vec{u}) = E\{\max_k [U^k + \epsilon^k] | \vec{U} = \vec{u}\}. \tag{33}$$

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

$$\frac{\mathcal{G}(u_*^0, \ldots, u_*^K + \eta) - \mathcal{G}(u_*^0, \ldots, u_*^K)}{\eta} = E\left[\frac{\Delta(\eta, \vec{u}_*, \vec{\epsilon})}{\eta}\right], \tag{34}$$

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

$$\Delta(\eta, \vec{u}_*, \vec{e}_*)/\eta = [(u_*^K + \eta + e^K) - (u_*^K + e^K)]/\eta = 1. \tag{35}$$

Thus,

$$\lim_{\eta\to 0} \frac{\Delta(\eta, \vec{u}_*, \vec{e})}{\eta} = 1. \tag{36}$$

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

$$\Delta(\eta, \vec{u}_*, \vec{e}_*)/\eta = [0]/\eta = 0. \tag{37}$$

Thus,

$$\lim_{\eta\to 0} \frac{\Delta(\eta, \vec{u}_*, \vec{e})}{\eta} = 0. \tag{38}$$

By Assumption 2.1(b), we have $\Pr(\epsilon^K + u_*^K = \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]) = 0$. Therefore, almost surely,

$$\lim_{\eta\to 0} \frac{\Delta(\eta, \vec{u}_*, \vec{\epsilon})}{\eta} = 1\{\epsilon^K + u_*^K > \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]\}. \tag{39}$$

Also, observe that

$$\left|\frac{\Delta(\eta, \vec{u}_*, \vec{\epsilon})}{\eta}\right| \leq \left|\frac{u_*^K + \eta + \epsilon^K - (u_*^K + \epsilon^K)}{\eta}\right| = 1 < \infty. \tag{40}$$

Thus, the bounded convergence theorem applies and yields

$$\lim_{\eta\to 0} E\left[\frac{\Delta(\eta, \vec{u}_*, \vec{\epsilon})}{\eta}\right] = E[1\{\epsilon^K + u_*^K > \max_{k=0,\ldots,K-1}[u_*^k + \epsilon^k]\}] = p^K(\vec{u}). \tag{41}$$

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

Proof. [Proof of Theorem 1] For both part (a) and part (b), the following fact is useful:

$$\mathcal{C} \subseteq H(\beta) := \{c \in \mathbb{R}^{d_x} : \beta' c \geq 0\}. \tag{42}$$

This is a fact because by the definition of $\mathcal{C}$, for every $c \in \mathcal{C}$, we have $\beta' c \geq 0$.

(a) Now we show the sufficiency. Assumption 2.3 implies that there exists a unique $b_* \in \mathbb{R}^{d_x}$ such that $\mathcal{C} = \{c \in \mathbb{R}^{d_x} : b_*' c \geq 0\}$. It suffices to show that $\beta = b_*$, or in other words, $\mathcal{C} = H(\beta)$. Suppose not. Then, there must exsits $c_* \in H(\beta)$ such that $c_* \notin \mathcal{C}$. In other words, c_* satisfies $b_*' c_* < 0$ and $\beta' c_* \geq 0$. Because $\beta \neq 0$ (by Assumption 2.2(a)), there exists a small purturbation ε to c_* such that $\beta'(c_* + \varepsilon) > 0$ and $b_*'(c_* + \varepsilon) < 0$. Let $c_{**} = -(c_* + \varepsilon)$. Then $\beta' c_{**} < 0$ and $b_*' c_{**} > 0$, or in other words, $c_{**} \in \mathcal{C}$ and $c_{**} \notin H(\beta)$. This contradicts (42). Therefore, $\mathcal{C} = H(\beta)$ and hence β is uniquely identified as the b_* that defines the half-space that $\mathcal{C}$ is.

Next we show the necessity. By definition, $\mathcal{C}$ is a pointed convex cone in $\mathbb{R}^{d_x}$. Every pointed convex cone is the intersection of all half-spaces (of the shape $H(b) := \{c \in \mathbb{R}^{d_x} : b' c \geq 0\}$ for some $b \in \mathbb{R}^{d_x}$) containing the cone. Suppose that Assumption 2.3 does not hold. Then $\mathcal{C} \neq H(\beta)$. Then there exists at least a $b_* \neq \beta$ such that $\mathcal{C} \subseteq H(\beta) \cap H(b_*)$. Because $\mathcal{A} \subseteq \mathcal{C}$, we have $\mathcal{A} \subseteq H(b_*)$. That is, the identifying inequalities (9) are satisfied with β replaced by b_* as well. Thus, β^* is not uniquely identified by those inequalities. This shows the necessity.

(b) It suffices to show that $\mathcal{C} = H(\beta)$ because $H(\beta)$ is a half-space. Let c_* be an arbitrary point in the *interior* of $H(\beta)$. By Assumption 2.4, it must be the case that there exists $\lambda_* > 0$, and $x_{1*}^1, x_{2*}^1 \in \mathcal{X}$ such that $c_* = \lambda_*(x_{2*}^1 - x_{1*}^1)$ and $\beta'(x_{2*}^1 - x_{1*}^1) > 0$.

Assumption 2.1(a) implies that $E(Y^1|X^1 = x^1) = \Pr(\epsilon^1 - \epsilon^0) > -\beta' x^1) = 1 - F_{\epsilon^1 - \epsilon^0}(-\beta' x^1)$. This and Assumption 2.1(c) together imply that $E(Y^1|X^1 = x^1)$ is strictly increasing in $\beta' x^1$. Thus, we have $E(Y^1|X^1 = x_{2*}^1) - E(Y^1|X^1 = x_{1*}^1) > 0$. Let b_* denote $E(Y^1|X^1 = x_{2*}^1) - E(Y^1|X^1 = x_{1*}^1)$. Then

$$c_* = b_*^{-1}\lambda_* \times (x_{2*}^1 - x_{1*}^1)[E(Y^1|X^1 = x_{2*}^1) - E(Y^1|X^1 = x_{1*}^1)]. \tag{43}$$

Because $b_*^{-1}\lambda_* > 0$ and because $(x_{2*}^1 - x_{1*}^1)[E(Y^1|X^1 = x_{2*}^1) - E(Y^1|X^1 = x_{1*}^1)] \in \mathcal{A}$, we have $c_* \in \mathcal{C}$. This shows that

$$H(\beta) \subseteq \mathcal{C}. \tag{44}$$

This combined with (42) implies that $\mathcal{C} = H(\beta)$.

Proof. [Proof of Theorem 3] Below we show that

$$\sup_{b=(b_1,\ldots,b_{d_x})' : \max_\ell |b_\ell| = 1} |Q(b) - Q_n(b)| \to_p 0. \tag{45}$$

Given (45), consider the following standard consistency derivation: for an arbitrary $\varepsilon > 0$,

$$\begin{aligned}
\Pr(\|\widehat{\beta} - \beta\| > \epsilon) &\leq \Pr(Q(\widehat{\beta}) - Q(\beta) > \delta(\varepsilon)) \\
&= \Pr(Q(\widehat{\beta}) - Q_n(\widehat{\beta}) + Q_n(\widehat{\beta}) - Q_n(\beta) + Q_n(\beta) - Q(\beta) > \delta(\varepsilon)) \\
&\leq \Pr(Q(\widehat{\beta}) - Q_n(\widehat{\beta}) + Q_n(\beta) - Q(\beta) > \delta(\varepsilon)) \\
&\leq \Pr\left(2 \sup_{\substack{b=(b_1,\ldots,b_{d_x})': \\ \max_\ell |b_\ell|=1}} |Q_n(b) - Q(b)| > \delta(\varepsilon)\right) \\
&\to 0, \qquad (46)
\end{aligned}$$

where the first inequality holds for some $\delta(\varepsilon) > 0$ by condition (i), the continuity of $Q(b)$ and the compactness of $\{b = (b_1, \ldots, b_{d_x})' : \max_\ell |b_\ell| = 1\}$, the second inequality holds because $Q_n(\widehat{\beta}) \leq Q_n(\beta)$, and the convergence holds by (45). This shows the theorem.

Now we show (45). First we show the stochastic equicontinuity of $Q_n(b)$. Consider the following derivation:

$$\begin{aligned}
&|Q_n(b) - Q_n(b^*)| \\
&\leq \max_{J=2,\ldots,\bar{J}} \max_{\substack{1\leq i_1,\ldots,i_J\leq n, \\ i_{J+1}=i_1}} \left| (b-b^*)' \frac{J^{-1}\sum_{j=1}^J \sum_{k=0}^K (X_{i_{j+1}}^k - X_{i_j}^k)\hat{p}^k(\vec{X}_{i_j})}{\max_{j=1,\ldots,J,k=0,\ldots,K} \|X_{i_j}^k\|} \right| \\
&\leq \max_{J=2,\ldots,\bar{J}} \max_{\substack{1\leq i_1,\ldots,i_J\leq n, \\ i_{J+1}=i_1}} \|b-b^*\| J^{-1} \sum_{j=1}^J \sum_{k=0}^K \frac{\|X_{i_{j+1}}^k - X_{i_j}^k\|}{\max_{j=1,\ldots,J,k=0,\ldots,K} \|X_{i_j}^k\|} \hat{p}^k(\vec{X}_{i_j}) \\
&\leq 2(K+1)\|b-b^*\|. \qquad (47)
\end{aligned}$$

Therefore,

$$\lim_{\delta\to 0} \sup_{b,b^*:\|b-b^*\|\leq\delta} |Q_n(b) - Q_n(b^*)| \leq \lim_{\delta\to 0} 2(K+1)\delta = 0 \ a.s. \qquad (48)$$

Given the stochastic equicontinuity (48) and the compactness of $\{b = (b_1, \ldots, b_{d_x})' : \max_\ell |b_\ell| = 1\}$, to show (45), it suffices to show that for all $b = (b_1, \ldots, b_{d_x})' : \max_\ell |b_\ell| = 1$, we have

$$Q_n(b) \to_p Q(b). \qquad (49)$$

For this purpose, we let $p^k(\vec{x}) = E(Y^k|\vec{X} = \vec{x})$, and let

$$\widetilde{Q}_n(b) = \max_{J=2,\ldots,\bar{J}} \max_{\substack{1\leq i_1,\ldots,i_J\leq n, \\ i_{J+1}=i_1}} \left[\frac{J^{-1}\sum_{j=1}^J \sum_{k=0}^K (b'X_{i_{j+1}}^k - b'X_{i_j}^k)p^k(\vec{X}_{i_j})}{\max_{j=1,\ldots,J,k=0,\ldots,K} \|X_{i_j}^k\|} \right]_+ \qquad (50)$$

By condition (iii), we have $\widetilde{Q}_n(b) \to_p Q(b)$. Now we only need to show that $|\widetilde{Q}_n(b) - Q_n(b)| \to_p 0$.

Consider the following derivation:

$$
\begin{aligned}
&|\widetilde{Q}_n(b) - Q_n(b)| \\
&\leq \max_{J=2,\ldots,\bar{J}} \max_{\substack{1\leq i_1,\ldots,i_J\leq n, \\ i_{J+1}=i_1}} \left[\frac{J^{-1}\sum_{j=1}^{J}\sum_{k=0}^{K} |b'X_{i_{j+1}}^k - b'X_{i_j}^k| \times |p^k(\vec{X}_{i_j}) - \hat{p}^k(\vec{X}_{i_j})|}{\max_{j=1,\ldots,J,k=0,\ldots,K} \|X_{i_j}^k\|} \right] \\
&\leq \max_{\vec{x}\in\vec{\mathcal{X}}} \left(2(K+1)\|b\|\right) |p^k(\vec{x}) - \hat{p}^k(\vec{x})| \\
&\to_p 0, \qquad (51)
\end{aligned}
$$

where the convergence holds by condition (ii).