# Robust Semiparametric Estimation in Panel Multinomial Choice Models*†

Wayne Yuan Gao‡ and Ming Li§

March 27, 2021

**Abstract**

This paper proposes a robust method for semiparametric identification and estimation in panel multinomial choice models, where we allow for infinite-dimensional fixed effects that enter into consumer utilities in an additively nonseparable way, thus incorporating rich forms of unobserved heterogeneity. Our identification strategy exploits multivariate monotonicity in parametric indexes, and uses the logical contraposition of an intertemporal inequality on choice probabilities to obtain identifying restrictions. We provide a consistent estimation procedure, and demonstrate the practical advantages of our method with simulations and an empirical illustration with the Nielsen data.

*We are grateful to Xiaohong Chen, Peter Phillips, and Phil Haile for their invaluable advice and encouragement. We thank Don Andrews, Isaiah Andrews, Tim Armstrong, Tim Christensen, Ben Connault, Bo Honoré, Joel Horowitz, Yuichi Kitamura, Patrick Kline, Lixiong Li, Charles Manski, Aviv Nevo, Matt Seo, Xiaoxia Shi, Frank Schorfheide, Ed Vytlacil, Rui Wang, Sheng Xu and seminar participants at Georgetown, UCSD, Berkeley, UCL, Northwestern, UW-Madison, UPenn, NYU, Princeton, HKU, CUHK, SMU, LSE, KU Leuven, Sciences Po, PSU, Microeconometrics Class of 2019 Conference (Duke) and 2020 Winter Meeting of the ES for helpful comments. All remaining errors are ours.

†Researchers own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

‡Gao: University of Pennsylvania, 133 S 36th St, Philadelphia, PA 19104. waynegao@upenn.edu.

§Li: Yale University, 28 Hillhouse Ave., New Haven, CT 06511. ming.li@yale.edu.

# 1 Introduction

The prevalence of heterogeneity and its importance in economic research are now well recognized. As pointed out by Heckman (2001), one of the most important discoveries in microeconometrics is the pervasiveness of diversity in economic behavior, which in turn has profound theoretical and practical implications. Browning and Carro (2007) survey the treatment of heterogeneity in applied microeconometrics, and find that "there is usually much more heterogeneity than researchers allow for", arguing that it is important yet difficult to accommodate heterogeneity in satisfactory ways. Moreover, the increasing availability of vast digital databases in this so-called "Big Data Era" brings about new challenges as well as opportunities for the treatment and understanding of heterogeneity (Fan, Han, and Liu, 2014).

More concretely, in analyzing consumer choices, a topic of wide theoretical and practical interest in microeconometrics, there might be rich forms of unobserved heterogeneity in consumer and product characteristics that influence choice behavior in significant yet complex ways. For example, it has long been recognized that brand loyalty is an important factor in determining choices of consumer products (Howard and Sheth, 1969), and research by Reichheld and Schefter (2000) along with their colleagues from Bain & Company, a leading management consulting firm, finds that brand loyalty is becoming even more important for online businesses. However, in modeling of consumer behavior it is very difficult (Luarn and Lin, 2003) to incorporate brand loyalty, a potentially complicated object that is clearly heterogeneous, hard to measure, and often unobserved in data. Besides brand loyalty, there may also be other forms of unobserved heterogeneity, such as subtle flavors and packaging designs, that may influence our choices of consumer products in everyday life. It is neither theoretically nor empirically clear whether all such complicated forms of unobserved heterogeneity can be fully captured by scalar-valued fixed effects in fully additive models, as often found in the literature.

Given these motivations, this paper proposes a simple and robust method for semiparametric identification and estimation in a panel multinomial choice model, where we allow for infinite-dimensional (functional) fixed effects that enter into consumer utilities in an additively nonseparable and thus fully flexible way, incorporating rich forms of unobserved heterogeneity. Our identification strategy exploits multivariate monotonicity in its contrapositive form, which provides powerful leverage for convert-

ing observable events into identifying restrictions under lack of additive separability. We provide consistent estimators based on our identification strategy, together with a computational algorithm implemented in a spherical-coordinate reparameterization that brings about a combination of topological, geometric and arithmetic advantages. A simulation study and an empirical illustration using the Nielsen data on popcorn sales are conducted to analyze the finite-sample performance of our estimation method and demonstrate the adequacy of our computational procedure for practical implementation.

We consider the following panel multinomial choice model in a short-panel setting:

$$y_{ijt} = \mathbb{1}\left\{u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) \geq \max_{k\in\{1,\ldots,J\}} u\left(X_{ikt}'\beta_0,\, A_{ik},\, \epsilon_{ikt}\right)\right\}$$

where agent $i$'s utility from a candidate product $j$ at time $t$, represented by $u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)$, is taken to be a function of three components. The first is a linear index $X_{ijt}'\beta_0$ of observable characteristics $X_{ijt}$, which contains a finite-dimensional parameter of interest $\beta_0$ we will identify and estimate. The second term $A_{ij}$ is an infinite-dimensional fixed effect matrix that can be heterogeneous across each agent-product combination. We emphasie that $X_{ijt}$ and $A_{ij}$ can be arbitrarily correlated, thus the model considered here falls in the broadly defined category of fixed effect panel models (Hsiao, 2014). The last term $\epsilon_{ijt}$ is an idiosyncratic time-varying error term of arbitrary dimensions. The three components are then aggregated by an unknown utility function $u$ in an additively nonseparable way, with the only restriction being that each agent's utility $u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)$ is *increasing* in its first argument, i.e., the linear index of observable characteristics $X_{ijt}'\beta_0$. Each agent then chooses a certain product in a given time period, represented by $y_{ijt} = 1$, if and only if this product gives him the highest utility among all available products.

The infinite-dimensionality of the terms $u$, $A_{ij}$ and $\epsilon_{ij}$ and the additive nonseparability in their interactions jointly produce rich forms of unobserved heterogeneity. Across each agent-product combination $ij$, we are effectively allowing for flexible variations in agent utilities as functions of the index $X_{ijt}'\beta_0$, which serve as nonparametric proxies for the effects of complicated unobserved factors that influence choice behavior, including brand loyalty, subtle flavors and packaging designs as discussed earlier. Moreover, unrestricted heterogeneity in the distribution of the error term $\epsilon_{ijt}$ is ac-

commodated, allowing for in particular heteroskedasticity in agent random utilities .

The generality of our setup encompasses many semiparametric (or parametric) panel multinomial choice models with scalar-valued fixed effects, scalar-valued error terms and various degrees of additive separability in the previous literature, including the following standard formulation:

$$y_{ijt} = \mathbb{1}\left\{X_{ijt}^{'}\beta_0 + A_{ij} + \epsilon_{ijt} \geq \max_{k\in\{1,\ldots,J\}}\left(X_{ikt}^{'}\beta_0 + A_{ik} + \epsilon_{ikt}\right)\right\}.$$

Relatively speaking, in this paper we are able to accommodate the infinite dimensionality of unobserved heterogeneity and the lack of additive separability in agent utility functions, under a standard time homogeneity assumption on the idiosyncratic error term that is widely adopted in the related literature.

Our key identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form. The idea is very simple and intuitive, and can be loosely described as the following: whenever we observe a *strict increase* in the choice probabilities of a specific product from one period to another, by logical contraposition it *cannot* be possible that this product becomes *worse* while all other products become *better* over the two periods. More formally, we show that a certain configuration of conditional choice probabilities satisfies the standard notion of weak multivariate monotonicity in all product indexes, which is naturally induced by the multinomial nature of our model and the monotonicity of each agent's utility function in each product's index. Then, we construct a collection of observable inequalities on conditional choice probabilities based on intertemporal comparison and cross-sectional aggregation, which preserves weak monotonicity in the index structure. Finally, we simply take a logical contraposition of the inequality on conditional choice probabilities, and obtain an identifying restriction on the index values free of all infinite-dimensional nuisance parameters, with which we construct a population criterion function that is guaranteed to be minimized at the true parameter value. The validity of this idea relies only on monotonicity in an index structure, and therefore it may have wider applicability beyond multinomial choice models.

Based on our identification result, we provide consistent set (or point) estimators, together with a computational algorithm adapted to the technical niceties and challenges of our framework. Specifically, our estimator can be computed through

a two-stage procedure. The first stage takes the form of a standard nonparametric regression, where we nonparametrically estimate a collection of intertemporal differences in conditional choice probabilities, using a machine learning algorithm based on artificial neural networks. In the second stage, we numerically minimize our sample criterion function, constructed as the sample analog of our population criterion function with the first-stage nonparametric estimates plugged in. A highlight of our estimation and computation procedure is the adoption of a spherical-coordinate reparameterization of our criterion functions in terms of *angles*, which enables us to exploit a combination of topological, geometric and computational advantages.

A simulation study is conducted to analyze the finite-sample performance of our method and the adequacy of our computational procedure for practical implementation. We investigate the performances of the first-stage and the final estimators under different model configurations, and show how the results vary with the sizes and dimensions of data. We also compare the performances of our estimator under set identification and point identification, and demonstrate the informativeness of our set estimator under the lack of point identification.

An empirical illustration of our procedure is also provided, where we use the Nielsen data on popcorn sales in the United States to explore the effects of marketing promotion effects. The results show that our procedure produces estimates that conform well with economic intuition. For example, we find that special in-store displays boost sales not only through a direct promotion effect but also through the attenuation of consumer price sensitivity, a result that cannot be produced by other methods based on additive separability. Intuitively, marketing managers are more likely to promote products that they know consumers are more price and promotion sensitive to. Hence, the average effective price sensitivity of promoted products tend to be larger than those not promoted due to the selection effect. Given the nonadditive nature of such selection effects, estimators based on additive separability will be biased. In contrast, our method is robust to such confounding effects, thus producing more economically sensible estimates.

As a further generalization, we discuss the wider applicability of our identification strategy beyond panel multinomial choice models, using an umbrella framework called *monotone multi-index* models. This framework captures the key ingredients of a large class of models, such as sample selection models and network formation models. In particular, we provide a specific illustration of a dyadic network formation model

under the setting of nontransferable utility, which naturally induces lack of additive separability in a micro-founded manner. The applicability of our current method, though with some nontrivial adaptions to the additional complications in network settings, is investigated in a companion paper by Gao, Li, and Xu (2020).

This paper builds upon and contributes to a large literature in econometrics on semiparametric (and parametric) discrete choice models, dating back to McFadden (1974) and Manski (1975), and more specifically a recent branch of research that focuses on panel multinomial choice models.

Our work is most closely related to the work by Pakes and Porter (2016), who also exploit weak monotonicity and time homogeneity. Our current paper adopts a similar approach that heavily exploits monotonicity, but does not restrict the effect of unobserved heterogeneity as a scalar index that is additively separable from the scalar index of observable characteristics. Hence, it is no longer feasible in our model to directly calculate the differences between the indexes of observable characteristics as in Pakes and Porter (2016).

Another related paper is Shi, Shum, and Song (2018), who propose a novel approach that exploits cyclical monotonicity of *vector*-valued functions in a fully additive panel multinomial choice model, where scalar-valued fixed effects are differenced out through "cyclical summation". Khan, Ouyang, and Tamer (2019) consider a similar additive multinomial choice model, but utilize the subsample of observations with time-invariant covariates along *all products but one* so as to leverage monotonicity in a single linear index for the construction of a rank-based estimator a la Manski (1987). Relatedly, the earlier work by Honoré and Kyriazidou (2000) also exploits monotonicity in a single index when certain covariates across two periods are equal in a dynamic panel setting. Another recent paper by Chernozhukov, Fernández-Val, and Newey (2019) studies a nonseparable multinomial choice model with bounded derivatives, and demonstrates semiparametric identification in a specialized panel setting with an additive effect under an "on-the-diagonal" restriction (i.e., when covariates at two different time periods coincide). Our method is significantly different from and thus complementary to those proposed in these afore-cited papers.

At a more general level, our work can be related to and compared to semiparametric methods of identification and estimation in *monotone single-index models*. A related class of estimators that leverage univariate monotonicity, known as *maximum*

*score* or *rank-order estimator*s, date back to a series of important contributions by Manski (1975, 1985, 1987), and are further investigated in Han (1987), Horowitz (1992), Abrevaya (2000), Honoré and Lewbel (2002) and Fox (2007). Despite the similarity in the reliance on monotonicity, the multinomial or *multi-index* nature of our current model induces a key difference from the single-index setting, leading to a significantly different method of estimation relative to rank-order estimators.

Finally, our model and method are complementary to another class of models that fall into the framework of *invertible multi-index models*. The celebrated paper by Berry, Levinsohn, and Pakes (1995) first utilizes the invertibility of the market share function to obtain a vector of unknown indexes, which is investigated more generally by Berry, Gandhi, and Haile (2013) and Berry and Haile (2014). Outside the context of demand estimation, a recent paper by Ahn, Ichimura, Powell, and Ruud (2018) provides a high-level treatment of multi-index models based on invertibility. In comparison, our paper does not involve invertibility, but relies on monotonicity.

The rest of this paper is organized as follows. Section 2 introduces our main model specifications and assumptions. Section 3 presents our key identification strategy. In Section 4 we provide consistent estimators along with a computational procedure to implement it. Section 5 and Section 6 contain a simulation study and an empirical illustration with the Nielsen data. Section 7 discusses the generalization of our method to monotone multi-index models, and finally we conclude with Section 8.

# 2 Panel Multinomial Choice Model

## 2.1 Model Setup

In this section we present a semiparametric panel multinomial choice model featured by infinite-dimensional unobserved heterogeneity and flexible forms of nonseparability, which we will use as the main model to illustrate our identification and estimation method. See Section 7 for a more general discussion about the wide applicability of our proposed methods.

Specifically, we consider the following discrete choice model, which states that agent $i$ chooses product $j$ at time $t$ if and only if $i$ prefers product $j$ to all other

alternatives at time $t$:

$$y_{ijt} = \mathbb{1}\left\{u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) \geq \max_{k\in\{0,1,\ldots,J\}} u\left(X_{ikt}'\beta_0,\, A_{ik},\, \epsilon_{ikt}\right)\right\} \tag{1}$$

where:

- $i \in \{1, ...N\}$ denotes $N$ decision makers, or simply *agents*.
- $j \in \{0, 1..., J\}$ denotes $J + 1$ choice alternatives, with $J$ *products* indexed by $1, ..., J$ and an *outside option* denoted by 0.
- $t \in \{1, ..., T\}$ denotes $T \geq 2$ different time periods.
- $X_{ijt}$ is $\mathbb{R}^D$-valued vector of observable characteristics specific to each agent-product-time tuple $ijt$. This could include, for example, buyer characteristics such as income level, product characteristics such as price and promotion status, as well as interaction and higher-order terms of those characteristics.
- $y_{ijt}$ is an observable binary variable, with $y_{ijt} = 1$ indicating that buyer $i$ chooses products $j$ at time $t$ and $y_{ijt} = 0$ indicating otherwise.
- $\beta_0 \in \mathbb{R}^D$ is a finite-dimensional unknown parameter of interest. We will repeatedly refer to the term $\delta_{ijt} := X_{ijt}'\beta_0$ as the ($ijt$-specific) *index* throughout this paper, which is intended to capture how the observable characteristics $X_{ijt}$ influence agent $i$'s choice of $j$ at $t$, *ceteris paribus*. Further discussion on the index is offered later.
- $A_{ij}$ represents an $ij$-specific time-invariant unobserved heterogeneity term of arbitrary dimensions, which we will refer to as the ($ij$-specific) *fixed effect*.
- $\epsilon_{ijt}$ is an $ijt$-specific unobserved error term of arbitrary dimensions, which captures time-idiosyncratic utility shocks to product $j$ for agent $i$ at time $t$.
- $u$ is an unknown function, interpreted as a *utility function* that aggregates the parametric index $X_{ijt}'\beta_0$, the fixed effect $A_{ij}$ and the error term $\epsilon_{ijt}$ into a scalar representing agent $i$'s utility from choosing product $j$ at time $t$.

We now provide some further clarifications and explanations for model (1).

We begin with a brief comparison that highlights the differences between our current model (1) to other models studied in several closely related papers on panel multinomial choice models. Notice first that model (1) includes as a special case the standard panel multinomial choice model under full additivity and scalar-valued unobserved heterogeneity:

$$y_{ijt} = \mathbb{1}\left\{X_{ijt}'\beta_0 + A_{ij} + \epsilon_{ijt} \geq \max_{k\in\{1,\ldots,J\}} X_{ikt}'\beta_0 + A_{ik} + \epsilon_{ikt}\right\}. \tag{2}$$

Such models have been studied in recent work by Khan, Ouyang, and Tamer (2019) and Shi, Shum, and Song (2018) with different methods of identification and estimation. In another recent paper by Pakes and Porter (2016), they investigate a generalized version of (2) in the following form:

$$y_{ijt} = \mathbb{1}\left\{g_j\left(X_{ijt}, \beta_0\right) + f_j\left(A_{ij}, \epsilon_{ijt}\right) \geq \max_{k\in\{1,\ldots,J\}} g_k\left(X_{ikt}, \beta_0\right) + f_k\left(A_{ik}, \epsilon_{ikt}\right)\right\}, \tag{3}$$

where the function $g_j$ produces a potentially nonlinear parametric index and $f_j$ aggregates fixed effects and idiosyncratic errors into a scalar value in a nonseparable way, while additive separability between the observable covariate index $g_j\left(X_{ijt}, \beta_0\right)$ and the unobserved heterogeneity index $f_j\left(A_{ij}, \epsilon_{ijt}\right)$ is still maintained. Moreover, although the dimensions of $(A_{ij}, \epsilon_{ijt})$ are not restricted in Pakes and Porter (2016), their overall effect is taken to be represented by a scalar value, $f_j\left(A_{ij}, \epsilon_{ijt}\right)$. We reiterate that our model (1) not only incorporates infinite-dimensionality in unobserved heterogeneity as captured by $A_{ij}$ and $\epsilon_{ijt}$, but also allows such heterogeneity to enter into agent utility functions in a fully *nonseparable* way.

The combination of infinite dimensionality and nonseparability jointly produces rich forms of heterogeneity in agent utility functions. Particularly, nonseparability translates into unrestricted flexibility regarding the ways in which the nonparametric fixed effect $A_{ij}$ may enter into the utility function $u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)$. In fact, we could equivalently suppress the notation $A_{ij}$ and instead write the utility function $u$ to be $ij$-specific,[1] i.e., $u_{ij}\left(X_{ijt}'\beta_0, \epsilon_{ijt}\right) \equiv u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)$. Written in this form, our formulation allows for flexible time-invariant heterogeneity in how the index $X_{ijt}'\beta_0$

[1]This reformulation, however, will introduce randomness to the utility function $u_{ij}$ when we consider the sampling process and assume cross-sectional random sampling later. Hence, to fully separate random elements from nonrandom ones, and to explicitly emphasize the dependence on $A_{ij}$, we will retain the notations of model (1) unless explicitly stated otherwise.

affects agent $i$'s utility from product $j$. In other words, given a fixed value of the index $\overline{\delta}$, the utility $u_{ij}\left(\overline{\delta}, \epsilon_{ijt}\right)$ can vary across each agent-product pair in totally unrestricted ways. Such heterogeneity can be induced by a plethora of complicated factors, such as subtle flavors, styles of design and social perceptions, the effects of which may be highly subjective on an individual basis. Some people may have a strong preference for Coca Cola over Pepsi or vice versa, while there might not exist any objective measure of flavor to assess, or even to describe, the subtle differences between the two popular soft drinks. Car shoppers may have heterogeneous tastes over engineering and design features in terms of safety, reliability, comfort, sportiness or luxury, while leading car manufacturers are often famous for their unique blends of features along these various dimensions, therefore appealing to different groups of customers to different extents. Beyond these examples, our formulation nests in itself arbitrary dimensions of agent-product specific heterogeneity that are time invariant.

It should be pointed out in particular that the fixed effect $A_{ij}$ effectively incorporates unobserved variations in the distributions of error terms $\epsilon_{ijt}$. For example, if we assume that $\epsilon_{ijt}$ is real-valued and follows a time-invariant distribution with a cumulative distribution function (CDF) $F_{ij}$, then the whole function $F_{ij}$ can be readily incorporated as part of the fixed effect $A_{ij}$, which may lie in a vector of infinite-dimensional functions. The CDF $F_{ij}$ absorbs a form of *heteroskedasticity* specific to each agent-product pair, and our method will be robust against such forms of heterogeneity in error distributions without the need to explicitly specify $F_{ij}$.

On a technical note, we now briefly discuss how the potential concern of tie-breaking can be handled in our framework. In cases where ties occur with nonzero probabilities, one popular approach in the literature is to incorporate a random tie-breaking process, modeled as a (potentially unknown) selection probability distribution among ties. The conceptual idea underlying this approach is to recognize the incompleteness of the model with respect to the determination of choice behaviors, and use an ad hoc selection probability to capture the effects of all unmodeled randomness. When we move from the scalar additive model (2) to model (1), rich forms of unmodeled randomness under (2) are automatically absorbed into the infinite-dimensional error term $\epsilon_{ijt}$, which nests in itself all possible latent variables that affect utilities in some appropriate yet unspecified ways.[2] As a result, the assumption

[2]It should be pointed out that the standard ad hoc approach, using selection probabilities among ties, and our current approach, where latent variables are explicitly modeled by the infinite-

that ties occur with zero probabilities is effectively a much weaker restriction under our current model (1) than under model (2).

The flexibility induced by nonseparability and infinite-dimensionality comes with the consequent analytical challenges to handle them. Various traditional techniques in the style of *differencing* based on additivity no longer work in our current model. For example, the recent method based on cyclical monotonicity proposed by Shi, Shum, and Song (2018) requires additivity to sum along a cycle of comparisons and cancel out the scalar-valued fixed effects via this summation, which becomes infeasible under nonseparability in our model (1). To confront the challenges induced by such nonseparability, we instead exploit a standard shape restriction, or more specifically, *monotonicity*, which captures a general commonality shared by many additive models but on its own does not involve additivity at all.

## 2.2 Key Assumptions

We now continue with a list of key assumptions required for our subsequent analysis, and discuss these assumptions in relation to model (1). To economize on notation, we will from now on frequently refer to the collection of variables concatenated along product and time dimensions: $\mathbf{X}_{it} := (X_{ijt})_{j=1}^{J}$, $\mathbf{X}_i = (\mathbf{X}_{it})_{t=1}^{T}$, $\mathbf{A}_i := (A_{ij})_{j=1}^{J}$, $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^{J}$ and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{it})_{t=1}^{T}$. Recall that we defined $\delta_{ijt} := X_{ijt}'\beta_0$. The first assumption below imposes a monotonicity restriction on the utility function.

**Assumption 1** (Monotonicity in the Index)**.** *$u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$ is weakly increasing in the index $\delta_{ijt}$, for every realization of $(A_{ij}, \epsilon_{ijt})$.*

It should first be clarified that the substantive part of Assumption 1 is the restriction of monotonicity in the index, while increasingness is without loss of generality given that the index $\delta_{ijt} = X_{ijt}'\beta_0$ contains an unknown parameter with unrestricted signs. Moreover, the monotonicity restriction is imposed on the index $\delta_{ijt}$, but not directly on any specific observable characteristics in $X_{ijt}$: quadratic or higher-order polynomial terms as well as other nonlinear or non-monotone functions of observable characteristics may be included in $X_{ijt}$ whenever appropriate.

dimensional error $\epsilon_{ijt}$, are two distinct approaches, neither of which includes the other as a special case. The key distinction comes from the *lexicographic* nature of the selection-probability approach, which cannot be fully represented by utility functions. It might be debatable whether the lexicographic structure is more conceptually justifiable or practically relevant, but we refrain from further discussion on this topic, as it is tangential to the main focus of this paper.

Assumption 1 not only serves as a key restriction that will be heavily leveraged upon by our subsequent identification and estimation method, but may also be regarded as an integral part of our semiparametric model: monotonicity endows the index $\delta_{ijt}$ with an interpretation as an objective summary statistic for the direct effect of observable covariates on agent utilities. In other words, $\delta_{ijt}$ may be considered as a quality measure of the match between agent $i$ and product $j$ based on their observable characteristics at time $t$, inducing a consequent interpretation of the parameter $\beta_0$ as representing how a certain change in a linear combination of observable characteristics may increase utilities for *all* agents from a certain product $j$, *ceteris paribus*.

Given the parametric index structure $\delta_{ijt} = X_{ijt}'\beta_0$, monotonicity itself seems a rather weak assumption widely satisfied in a large class of models. In many additive models where a parametric index in the style of $X_{ijt}'\beta_0$ is added to other components of the model, Assumption 1 could be trivially satisfied by construction, such as the standard panel multinomial choice model (2). In Section 7, we provide more examples of parametric and semiparametric models featured by monotonicity in an index structure beyond the multinomial choice setting.

**Assumption 2** (Cross-Sectional Random Sampling)**.** $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \boldsymbol{\epsilon}_i)$ *is i.i.d. across* $i \in \{1, ..., N\}$ *with* $N \to \infty$.

Assumption 2 is a standard assumption on random sampling.[3] In particular, we only require a *short panel*, where we focus on cross-sectional asymptotics with the number of agents getting large ($N \to \infty$) but the number of time periods $T$ held fixed.

**Assumption 3** (Conditional Time Homogeneity of Errors)**.** *The conditional distribution of* $\boldsymbol{\epsilon}_{it}$ *given* $(\mathbf{X}_i, \mathbf{A}_i)$ *is stationary over time* $t$*, i.e.,* $\boldsymbol{\epsilon}_{it} | (\mathbf{X}_i, \mathbf{A}_i) \sim \mathbb{P}(\cdot | \mathbf{A}_i)$.

Finally, we impose a conditional time homogeneity assumption on the idiosyncratic shocks. Assumption 3 is strictly stronger than necessary for our purpose, but leads to easier notations afterwards for clearer illustration of our key method. Alternatively, we could impose the following weaker version:

**Assumption 3'** (Pairwise Time Homogeneity of Errors)**.** *The marginal distributions of* $\boldsymbol{\epsilon}_{it}$ *and* $\boldsymbol{\epsilon}_{is}$ *conditional on* $(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i)$ *are the same across any pair of periods* $t \neq s \in \{1, ..., T\}$*, i.e.,* $\boldsymbol{\epsilon}_{it} | (\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i) \sim \boldsymbol{\epsilon}_{is} | (\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i)$.

[3] It is worth noting that so far we have not made any explicit restriction on the structure of the spaces on which the arbitrary dimensional random elements $\mathbf{A}_i$ and $\boldsymbol{\epsilon}_i$ are defined, but implicit in our specification as well as Assumption 2 is the requirement that $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \epsilon_i)$ be well-defined as random elements (measurable functions) on a large enough probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Assumption 3', a multinomial extension of the group homogeneity assumption in Manski (1987), is also imposed in Pakes and Porter (2016) and Shi, Shum, and Song (2018), both containing further discussions about the interpretation, flexibility and restrictions associated with this assumption. Assumption 3' suffices for our subsequent analysis based on pairwise intertemporal comparisons, while allowing for some dependence of $\boldsymbol{\epsilon}_{it}$ on time-varying component of observable covariates $(\mathbf{X}_{it}, \mathbf{X}_{is})$. We demonstrate in Appendix B that our identification and estimation results carry over under Assumption 3', but until then we will work with the stronger Assumption 3 for notational simplicity.

It might be worth noting that Assumption 3 (or 3'), a statement conditioned on the arbitrarily dimensional fixed effect $\mathbf{A}_i$ in a fully flexible manner, automatically absorbs all possible *time-invariant* components in $\mathbf{X}_{it} = (X_{ijt})_{j=1}^{J}$ and $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^{J}$. As discussed earlier, long-term brand loyalty, potentially produced by a mixture of complicated factors such as design, style, flavor, consumer personality or social perception, is just one example that applied researchers have found to be important since long ago (Howard and Sheth, 1969) yet conceptually difficult to incorporate empirically (Luarn and Lin, 2003). Such factors are often hard, if not impossible, to measure quantitatively and therefore are largely unobserved, and it is neither theoretically nor empirically clear whether a single-dimensional scalar term is sufficient to capture the effects from such factors. In the meanwhile, completely ignoring these factors will likely create endogeneity issues in econometric analysis of consumer behaviors, and it might be hard to find proper instruments for every potentially relevant latent factor. Therefore, we believe that our main model along with the assumptions above, admittedly with its own restriction to the fixed-effect specification, constitutes a step forward in the direction of accommodating more complex unobserved heterogeneity.

A noteworthy restriction of Assumption 3 lies in that it rules out random coefficients, a widely adopted modeling device proposed by Berry, Levinsohn, and Pakes (1995) to induce sophisticated substitution patterns among products with multi-dimensional characteristics space. However, the flexibility afforded by our general fixed effect specification can incorporate arbitrarily complicated substitution patterns with respect to *time-invariant* components of observed and unobserved product characteristics, by exploiting the panel structure of observable data along with the time homogeneity assumption (Assumption 3).[4] It is thus worth pointing out that our

[4]For example, consider the automobile industry analyzed in Berry, Levinsohn, and Pakes (1995,

current fixed-effect approach and the random-coefficient approach are two rather different methods: neither nests the other as a special case, and the two approaches may be more suitable for different sets of empirical applications. The random-coefficient approach using market share inversion, as developed by Berry, Levinsohn, and Pakes (1995), Berry, Gandhi, and Haile (2013) and Berry and Haile (2014), has already been widely used in various settings of demand analysis where time-varying (or market-varying) endogeneity is a major concern. Our infinite-dimensional fixed-effect approach based on weak monotonicity might be more suitable to panel-data settings where researchers are more interested in incorporating an arbitrarily complicated form of time-invariant heterogeneity across agent-product pairs.

Finally, as briefly discussed in Section 2.1 and formally stated in Assumption 3, the whole distribution of $\boldsymbol{\epsilon}_{it}$ can be indexed by the fixed effect $\mathbf{A}_i$. Furthermore, serial autocorrelation in $\boldsymbol{\epsilon}_{it}$ is not ruled out either, as Assumption 3 concerns only the marginal distributions of $\boldsymbol{\epsilon}_{it}$ in different periods.

We may now proceed to provide identification arguments for the leading parameter of interest, $\beta_0$, in Section 3 and construct estimators of $\beta_0$ in Section 4.

# 3 Identification Strategy

In this section, we present semiparametric identification results for model (2) under Assumptions 1-3. However, as will become clear later in this section, the underlying idea of our identification strategy applies more widely beyond panel multinomial choice models. See Section 7 for more details.

Our key identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form. As a reminder, we start with a standard definition of multivariate monotonicity, followed by a statement of its logical contraposition.

**Definition 1** (Multivariate Monotonicity)**.** A real-valued function $\psi : \mathbb{R}^J \to \mathbb{R}$ is said to be *weakly increasing* if, for any pair of vectors $\overline{\boldsymbol{\delta}}$ and $\underline{\boldsymbol{\delta}}$ in $\mathbb{R}^J$, if $\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j$ for

BLP), which can accommodate more reasonable substitution patterns between similar cars (say, among BMW, Mercedes and Yugo). In a panel setting with automobile data (recall that BLP considers a multi-market setting instead of a panel one), observable covariates such as "brands", "number of cylinders" and to a large extent "size" are time-invariant, and hence our current framework can accommodate flexible patterns of individual heterogeneity in the preference for "luxury", "6-cylinder" or "large" cars via the infinite-dimensional individual-and-product-specific fixed effect $A_{ij}$.

every $j = 1, ..., J$, then $\psi\left(\overline{\boldsymbol{\delta}}\right) \leq \psi\left(\underline{\boldsymbol{\delta}}\right)$.

*Remark* 1 (Logical Contraposition). The following is equivalent to Definition 1:

$$\psi\left(\overline{\boldsymbol{\delta}}\right) > \psi\left(\underline{\boldsymbol{\delta}}\right) \quad \Rightarrow \quad \text{NOT } \left\{\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j \text{ for all } j = 1, ..., J\right\}. \tag{4}$$

for any $\left(\overline{\boldsymbol{\delta}}, \underline{\boldsymbol{\delta}}\right)$, where "NOT" denotes the logical negation operator.

Our subsequent identification strategy will leverage heavily the simple contraposition of monotonicity (4), and our arguments proceed in three major steps. First, we define a multivariate monotone function in the form of conditional choice probabilities. Second, we construct an observable inequality based on the monotone function we define, effectively producing the left-hand side of (4). Finally, we use the contraposition of monotonicity to obtain the right-hand side of (4), which will translate into identifying restrictions on the parameter $\beta_0$ via the indexes $\boldsymbol{\delta}_{it} := (\delta_{ijt})_{j=1}^{J}$.

We now present our key identification strategy step by step. For the moment, we fix a particular product $j \in \{1, ..., J\}$, a pair of time periods $t \neq s \in \{1, ..., T\}$ and condition on a generic realization of the observable covariates in the two periods $t$ and $s$, i.e., $(\mathbf{X}_{it}, \mathbf{X}_{is}) = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \in Supp\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)$.

**Step 1: Construction of a monotone function**

For each individual $i$, consider $i$'s choice probability of $j$ given $(\mathbf{X}_{it}, \mathbf{A}_i)$:

$$\begin{aligned}
\mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right] &= \int \mathbb{1}\left\{u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \neq j} u\left(X_{ikt}'\beta_0, A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{it} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right) \\
&= \int \mathbb{1}\left\{u\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \neq j} u\left(\delta_{ikt}, A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{it} \middle| \mathbf{A}_i\right) \\
&=: \psi_j\left(\delta_{ijt}, \left(-\delta_{ikt}\right)_{k \neq j}, \mathbf{A}_i\right)
\end{aligned} \tag{5}$$

where the second equality follows from the index definition $\delta_{ijt} = X_{ijt}'\beta_0$ and Assumption 3 (Conditional Time Homogeneity of Errors), which enables us to write $\psi_j$ without the time subscript $t$. Clearly, the monotonicity of the utility function $u$ in the index argument $\delta_{ijt}$ (Assumption 1) translates into the multivariate monotonicity of the function $\psi_j$ in the vector of indexes $\left(\delta_{ijt}, \left(-\delta_{ikt}\right)_{k \neq j}\right)$[5]:

**Lemma 1.** *$\psi_j(\cdot, \mathbf{A}_i) : \mathbb{R}^J \to \mathbb{R}$ is weakly increasing, for any realized $\mathbf{A}_i$.*

[5]We flip the signs of $(\delta_{ikt})_{k \neq j}$ purely for the ease of exposition: as discussed earlier, it is the monotonicity, not the exact direction of monotonicity, that matters in our analysis.

In terms of economic interpretation, $\psi_j(\boldsymbol{\delta}_{it}, \mathbf{A}_i)$ summarizes each agent $i$'s conditional choice probability of product $j$ given $i$'s fixed effect $\mathbf{A}_i$ as a function of the index vector $\boldsymbol{\delta}_{it}$. Lemma 1 admits a simple interpretation: if a product $j$ becomes weakly better for agent $i$ (in terms of the index $\delta_{ijt}$), while all other products $k \neq j$ becomes weakly worse, then agent $i$'s choice probability of product $j$ should weakly increase.

However, as the realization of $\mathbf{A}_i$ is not observable, the conditional choice probability function $\psi_j(\cdot, \mathbf{A}_i)$ is not directly identified from data in the short-panel setting under consideration here. In the next step, we construct an observable quantity based on $\psi_j$ by averaging out $\mathbf{A}_i$.

**Step 2: Construction of an observable inequality**

Consider the following intertemporal difference in conditional choice probabilities:

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right] \tag{6}$$

which is by construction directly identified from data.

Write $\overline{\boldsymbol{\delta}} := \overline{\mathbf{X}}\beta_0 \equiv \left(\overline{X}_j'\beta_0\right)_{j=1}^{J}$ and similarly for $\underline{\boldsymbol{\delta}}$, and $\mathbf{X}_{i,ts} := (\mathbf{X}_{it}, \mathbf{X}_{is})$. The following lemma translates the monotonicity of $\psi_j\left(\overline{\boldsymbol{\delta}}, \mathbf{A}_i\right)$ in the index vector $\overline{\boldsymbol{\delta}}$ into a restriction on the sign of the observable quantity $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, effectively corresponding to an observable scalar inequality.

**Lemma 2.** $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j \implies \gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \leq 0$.

To see why Lemma 2 is true, rewrite $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ as

$$\begin{aligned}
\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) &= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right] \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{A}_i\right] - \mathbb{E}\left[y_{ijs} \middle| \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \\
&= \int \left[\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) - \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right).
\end{aligned}$$

Whenever $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j$, by Lemma 1 we have

$$\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) - \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) \leq 0$$

for *every* possible realization of $\mathbf{A}_i$. Consequently, the inequality will be preserved after integrating over the fixed effect $\mathbf{A}_i$ *cross-sectionally* with respect to the conditional

distribution $\mathbb{P}\left(\mathbf{A}_i | \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right)$, a potentially hugely complicated probability measure that we leave unspecified.

**Step 3: Derivation of the key identifying restriction**

We now take the logical contraposition of Lemma 2:

**Proposition 1** (Key Identifying Restriction)**.** *Under Assumptions 1, 2 and 3,*

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0 \;\Rightarrow\; \text{NOT } \left\{\left(\overline{X}_j - \underline{X}_j\right)' \beta_0 \leq 0 \text{ and } \left(\overline{X}_k - \underline{X}_k\right)' \beta_0 \geq 0 \; \forall k \neq j\right\} \tag{7}$$

Recall that $\delta_{ijt} = X_{ijt}'\beta_0$, so Proposition 1 follows immediately from Lemma 2 and defines an identifying restriction on $\beta_0$ that is free of all unknown nonparametric heterogeneity terms $u$, $\mathbf{A}$ and $\boldsymbol{\epsilon}$. Proposition 1 is also very intuitive: if we observe an intertemporal increase in the conditional choice probability of product $j$ from one period to another, it is impossible that product $j$'s index becomes worse, while all other products' indexes become better.

The simple idea behind Proposition 1 is to leverage the contraposition of monotonicity in the index vector, which, apart from its simplicity, brings about robustness against the rich built-in forms of unobserved heterogeneity along with nonseparability. As the validity of this idea relies only on monotonicity in an index structure, it is applicable more widely beyond the panel multinomial choice settings we are currently considering. See Section 7 for a general framework under which the contraposition of monotonicity may be utilized. In particular, in a companion paper (Gao, Li, and Xu, 2020), we adapt this idea to the additional complications induced in a network formation setting, where nonseparability arises naturally from nontransferable utilities.

We also note that the same idea can be readily extended to any nonempty subset of products, as summarized in the following corollary:

**Corollary 1.** *If* $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$ *for all* $j \in J_1 \subseteq \{0, 1, ..., J\}$*, it must NOT be that* $\left(\overline{X}_j - \underline{X}_j\right)' \beta_0 \leq 0$ *for all* $j \in J_1$ *while* $\left(\overline{X}_k - \underline{X}_k\right)' \beta_0 \geq 0$ *for all* $k \in J \backslash J_1$*.*

Intuitively, if we observe that the conditional choice probabilities of all products in $J_1$ strictly increase across two periods of time, it cannot be the case that the indexes of all products in $J_1$ have weakly worsened while the indices of all products outside $J_1$ have weakly improved. Li (2019) shows that, at least in the case of $T = 2$, the

collection of all identifying restrictions in Corollary 1 lead to *sharp* identification of $\beta_0$. That said, for the rest of the paper we will focus on the identifying restrictions in Proposition 1, while noting that all the analysis below can be readily adapted to incorporate the additional restrictions in Corollary 1.

**Formulation of Population Criterion Functions**

We now formulate a population criterion function based on Proposition 1. For every candidate parameter $\beta \in \mathbb{R}^D$, we represent in Boolean algebra the right hand side of (7) in Proposition 1 by

$$\lambda_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}; \beta\right) := \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k \neq j\}}\left(\overline{X}_k - \underline{X}_k\right)' \beta \leq 0\right\}, \tag{8}$$

where $(-1)^{\mathbb{1}\{k\neq j\}}$ takes the value $-1$ for $k \neq j$ and 1 for $k = j$. Therefore, Proposition 1 can be written algebraically as: $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$ implies $\lambda_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}; \beta_0\right) \equiv 0$ for any $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. We now define the following criterion function by taking a cross-sectional expectation over the random realization of $(\mathbf{X}_{it}, \mathbf{X}_{is})$:

$$Q_{j,t,s}(\beta) := \mathbb{E}\left[\mathbb{1}\left\{\gamma_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right) > 0\right\} \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right)\right], \tag{9}$$

which is clearly nonnegative and minimized to zero at the true parameter value $\beta_0$. Without normalization and further assumptions for point identification, there might be multiple values of $\beta_0$ that minimize $Q_{j,t,s}$ to zero.

More generally, fix any function $G : \mathbb{R} \to \mathbb{R}$ that is *one-sided sign preserving*, i.e., $G(z) > 0$ for $z > 0$ and $G(z) = 0$ for $z \leq 0$. For example, we can choose $G(z) = [z]_+$ where $[z]_+$ is the positive part function. Then, we define $Q^G_{j,t,s}$ as

$$Q^G_{j,t,s}(\beta) := \mathbb{E}\left[G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)\right) \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right)\right], \tag{10}$$

which is also minimized to zero at the true parameter value $\beta_0$. The sign-preserving function $G$, if also set to be monotone, continuous or bounded, serves as a *smoothing* function that helps with the finite-performance of our estimators. We will provide more discussions on function $G$ in the next section, when we construct estimators based on the sample analog of the population criterion function defined here. It is worth pointing out that this smoothing function $G$ is built into the *population* criterion

function as in (10), which is different from the usual technique where smoothing is only done in finite samples but not in the population. For notational simplicity, we suppress $G$ in $Q^G_{j,t,s}$ and simply write $Q_{j,t,s}$ throughout this paper.

So far we have focused on a fixed product $j$ and a fixed pair of periods $(t, s)$, but in practice we may utilize the information across all products and all pairs of periods by defining the aggregated criterion function:

$$Q(\beta) := \sum_{j=1}^{J} \sum_{t \neq s}^{T} Q_{j,t,s}(\beta), \quad \text{for any } \beta \in \mathbb{R}^D. \tag{11}$$

We summarize our main identification result in the following theorem.

**Theorem 1** (Set Identification)**.** *Under model* (1) *and Assumptions 1-3,*

$$\beta_0 \in B_0 := \left\{ \beta \in \mathbb{R}^D : Q(\beta) = 0 \right\}. \tag{12}$$

We will refer to $B_0$ as the *identified set*. In Appendix C, we provide sufficient conditions for point identification of $\beta_0$ up to scale normalization, with similar styles of assumptions imposed for point identification in the literature on maximum-score or rank-order estimation, dating back to Manski (1985), as well as in related work on panel multinomial choice models, such as Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2019).[6] However, since point identification, or lack thereof, is conceptually irrelevant to our key methodology, and as set identification and set estimation are becoming increasingly relevant in econometric theory as well as applied research, we will focus on set identification and estimation results in the main text, following a similar approach adopted by Manski (1975). Of course, whenever the additional assumptions for point identification are satisfied in data, the set estimator will shrink to a point asymptotically.

Our criterion function is constructed to be an aggregation of the identifying restrictions on $\beta_0$ in the form of Boolean variables across all $(j, t, s)$ in the data,

[6] It might be worth pointing out that the identification arguments in Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2019) feature conditioning on *equality* events in the form of $\left\{ \overline{X}_k - \underline{X}_k = \mathbf{0}, \text{ for all } k \neq j \right\}$, which essentially utilizes subsamples where observable covariates stay unchanged except for a single product $j$ across two periods. In contrast, our point identification argument, available in Appendix C, does not involve conditioning on *equalities*, but only *inequalities* that define (intersections of) half-spaces in the parameter space $\mathbb{R}^D$.

obtained via the logical contraposition of weak multivariate monotonicity whenever $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) > 0$ occurs. As $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) = -\gamma_{j,s,t}(\mathbf{X}_{is}, \mathbf{X}_{it})$, either $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) > 0$ or $\gamma_{j,s,t}(\mathbf{X}_{is}, \mathbf{X}_{it}) > 0$ occurs for each unordered pair of periods $\{t, s\}$, provided that there is nonzero intertemporal variation in the relevant conditional choice probabilities.

It is important to note that the stochastic relationship between the outcome variable $\mathbf{y}_i$ and the observable covariates $\mathbf{X}_i$ enters into our criterion function $Q$ only through the intertemporal differences in conditional choice probabilities as represented by the term $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is})$. As the randomness of $\mathbf{y}$ conditional on $\mathbf{X}$ is completely averaged out in $\gamma_{j,t,s}$, the only remaining form of randomness in our population criterion function is the random sampling of observable covariates $\mathbf{X}_i$, which no longer involves the outcome variable $\mathbf{y}_i$.

As a result, the systematic component of our population criterion function $Q_{j,t,s}$, as defined in (9) and (10), is nonstandard relative to usual forms of moment conditions as studied in the literature on extremum estimation. Specifically, in our criterion function the expectation (moment) operators show up twice, the first time in the definition of the conditional expectation $\gamma_{j,t,s}$ and the second time in the expectation over observable covariates $(\mathbf{X}_{it}, \mathbf{X}_{is})$. Moreover, the two expectation operators are separated by the *nonlinear* one-sided sign-preserving function $G$, so it is impossible to push inside the expectation operators via the law of iterated expectations.

Relative to the well-known maximum-score or rank-order criterion function as studied by Manski (1985, 1987) utilizing univariate monotonicity, the nonstandardness of our criterion function arises from a key difference of multivariate monotonicity from univariate monotonicity. To see this more clearly, consider the special case of a *single-index* setting ($J = 1$)[7], in which our population criterion function degenerates to the maximum-score or rank-order criterion function if we choose $G$ to be $G(z) = [z]_+$, suppress the product subscript $j$, and denote $X_t$ as the *vector* of observable covariates:

$$\begin{aligned} Q_{t,s}(\beta) + Q_{s,t}(\beta) = \mathbb{E}\left[[\gamma(X_t, X_s)]_+ \mathbb{1}\{(X_t - X_s)\beta \geq 0\}\right] \\ + \mathbb{E}\left[[\gamma(X_s, X_t)]_+ \mathbb{1}\{(X_s - X_t)\beta \geq 0\}\right] \end{aligned}$$

[7]This arises naturally in binomial choice models with the characteristics of the outside option set to be zero. In this case, even though there are nominally two choice alternatives, choice behavior is completely determined by a single index based on the characteristics of the non-default option.

$$= \mathbb{E}\left[\left(y_t - y_s\right) \operatorname{sgn}\left(\left(X_t - X_s\right)\beta\right)\right]. \tag{13}$$

The last line of (13) is the familiar maximum-score criterion function, constructed based on the following equivalence relationship induced by univariate monotonicity:

$$\left\{\gamma\left(X_t, X_s\right) > 0\right\} \;\Leftrightarrow\; \left\{\left(X_t - X_s\right)\beta > 0\right\}, \tag{14}$$

Such an equivalence relationship is a unique feature of the univariate setting, which can be derived as a special case of Proposition 1:

$$\left\{\gamma\left(X_t, X_s\right) > 0\right\} \Rightarrow \text{NOT}\left\{\left(X_t - X_s\right)\beta \leq 0\right\} \Leftrightarrow \left\{\left(X_t - X_s\right)\beta > 0\right\} \Rightarrow \left\{\gamma\left(X_t, X_s\right) \geq 0\right\},$$

which becomes (14) if the monotonicity of $\gamma$ is strict.

However, such equivalence relationships *cannot* be generalized to the multivariate setting with $J \geq 2$, as the right hand side of (7),

$$\text{NOT}\left\{\left(\overline{X}_j - \underline{X}_j\right)' \beta_0 \leq 0 \text{ and } \left(\overline{X}_k - \underline{X}_k\right)' \beta_0 \geq 0 \text{ for all } k \neq j\right\},$$

*does not* imply $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \geq 0$ in the converse direction. This breaks the equivalence built into the maximum-score criterion function. As a result, we can no longer aggregate $Q_{j,t,s}$ and $Q_{j,s,t}$ into a unified representation as in (13).

Hence, our population criterion function is a generalization of the maximum-score criterion functions to multi-index settings, where the lack of equivalence as described above leads to a key difference in the criterion functions, and consequently a different approach of estimation, which will be discussed in the next section.

# 4 Estimation and Computation

## 4.1 A Consistent Two-Step Estimator

We construct our estimator as a semiparametric two-step M-estimator.

The first stage of our procedure concerns with nonparametrically estimating the intertemporal differences in conditional choice probabilities of the following form

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right]$$

for all on-support realizations $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, all pairs of periods $(t, s)$ and all products $j$.[8]

Given the first-stage estimators $\hat{\gamma}_{j,t,s}$ and the smoothing function $G$, in the second stage we numerically compute minimizers of the sample criterion function,

$$\hat{Q}(\beta) := \sum_{j=1}^{J} \sum_{t \neq s}^{T} \hat{Q}_{j,t,s}(\beta),$$
$$\hat{Q}_{j,t,s}(\beta) := \frac{1}{N} \sum_{i=1}^{N} G\left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right) \lambda_j\left(\mathbf{X}_{i,ts}; \beta\right).$$

Observing that the scale of $\beta_0$ cannot be identified given that $\lambda_j\left(\mathbf{X}_{i,ts}; \beta\right)$ consists of indicator functions of the the form $\mathbb{1}\left\{\left(X_{ijt} - X_{ijs}\right)' \beta \geq 0\right\}$, we imposes the following scale normalization $\beta_0 \in \mathbb{S}^{D-1} := \left\{v \in \mathbb{R}^D : \|v\| = 1\right\}$. Following Chernozhukov, Hong, and Tamer (2007), we define the set estimator by

$$\hat{B}_{\hat{c}} := \left\{\beta \in \mathbb{S}^{D-1} : \ \hat{Q}(\beta) \leq \min_{\tilde{\beta} \in \mathbb{S}^{D-1}} \hat{Q}\left(\tilde{\beta}\right) + \hat{c}\right\} \tag{15}$$

with $\hat{c} := O_p\left(c_N \log N\right)$. We now introduce assumptions for the consistency of $\hat{B}_{\hat{c}}$.

**Assumption 4** (First-Stage Estimation)**.** *For any* $(j, t, s)$*:*

*(i)* $\gamma_{j,t,s} \in \Gamma$*, and* $\mathbb{P}\left(\hat{\gamma}_{j,t,s} \in \Gamma\right) \to 1$*, with* $\Gamma$ *being a* $\mathbb{P}$*-Donsker class of functions in* $L_2(\mathbf{X})$ *s.t.* $\sup_{\gamma_{j,t,s} \in \Gamma} \mathbb{E}\left|\gamma_{j,t,s}\right| < \infty$*;*

*(ii)* $\left\|\hat{\gamma}_{j,t,s} - \gamma_{j,t,s}\right\|_2 := \sqrt{\int \left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right) - \gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)^2 \mathrm{d}\mathbb{P}\left(\mathbf{X}_{i,ts}\right)} = O_p\left(c_N\right)$ with $c_N \searrow 0$ as $N \to \infty$.

Through Assumption 4 we take as given the large set of theoretical results on non-parametric regression in the literature. Many kernel-based and sieve-based methods have been developed with different properties demonstrated under various sets of conditions. See Wasserman (2006) and Chen (2007) for more comprehensive surveys.

[8] In practice, we only need to estimate $\gamma_{j,t,s}$ for $(J-1)$ products and $\frac{1}{2}T(T-1)$ *ordered* pairs of periods. The former is because conditional choice probabilities must sum to one across all $J$ products, so we may easily compute the estimator for the last product from the other $(J-1)$ estimates: $\gamma_{J,t,s} = 1 - \sum_{j=1}^{J-1} \gamma_{j,t,s}$. The latter is because $\gamma_{j,t,s} = -\gamma_{j,s,t}$ by construction, so we may estimate it for either $(t, s)$ or $(s, t)$. Notice, however, that each ordered pair $(t, s)$ or $(s, t)$ provides complementary identifying information, as $\lambda\left(\mathbf{X}_{i,ts}; \beta\right)$ and $\lambda\left(\mathbf{X}_{i,st}; \beta\right)$ do not admit such kind of deterministic relationship.

**Assumption 5** (Nice Smoothing Function)**.** *The one-sided sign-preserving function* $G : \mathbb{R} \to \mathbb{R}_+$ *is Lipschitz continuous with a finite Lipschitz constant.*

Assumption 5 is not necessary for consistency per se given that our identification result is valid with any choice of the one-sided sign-preserving function $G$, nevertheless we take $G$ to be Lipschitz so as to simplify the proof.

To state the next assumption, we decompose each row (corresponding to each product) of $\overline{\mathbf{X}} - \underline{\mathbf{X}}$ as the product of its norm and its *direction*, i.e., $\overline{X}_j - \underline{X}_j \equiv r_j\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) \cdot v_j\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right)$, where $r_j\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := \left\|\overline{X}_j - \underline{X}_j\right\|$, and $v_j\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := \left(\overline{X}_j - \underline{X}_j\right) / \left\|\overline{X}_j - \underline{X}_j\right\|$ if $\overline{X}_j \neq \underline{X}_j$ while $v_j\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := \mathbf{0}$ if $\overline{X}_j = \underline{X}_j$.

**Assumption 6** (Continuous Distribution of Directions)**.** *The marginal distribution of* $v_j\left(\mathbf{X}_{it} - \mathbf{X}_{is}\right)$ *has no mass point except possibly at* $\mathbf{0}$ *for each* $(j, t, s)$.

Assumption 6 is a technical assumption that ensures the continuity of the population criterion function. We note that Assumption 6 is fairly weak: it essentially requires that the *directions* of intertemporal differences in observable characteristics are continuously distributed on their own supports. In particular, this allows all but one dimensions of observable characteristics to be discrete.

With the above assumptions, we now establish the consistency of the set estimator $\hat{B}_{\hat{c}}$ based on Chernozhukov, Hong, and Tamer (2007).

**Theorem 2** (Consistency)**.** *Under Assumptions 1-6, the set estimator* $\hat{B}_{\hat{c}}$ *is consistent in Hausdorff distance:* $d_H\left(\hat{B}_{\hat{c}}, B_0\right) = o_p(1)$*, where*

$$d_H\left(\hat{B}_{\hat{c}}, B_0\right) := \max\left\{\sup_{\beta \in \hat{B}_{\hat{c}}} \inf_{\tilde{\beta} \in B_0} \left\|\beta - \tilde{\beta}\right\|, \sup_{\beta \in B_0} \inf_{\tilde{\beta} \in \hat{B}_{\hat{c}}} \left\|\beta - \tilde{\beta}\right\|\right\}.$$

*Furthermore, if* $\beta_0$ *is point-identified on* $\mathbb{S}^{D-1}$, $\left\|\hat{\beta} - \beta_0\right\| = o_p(1)$ *for any* $\hat{\beta} \in \hat{B} := \arg\min_{\tilde{\beta} \in \mathbb{S}^{D-1}} \hat{Q}\left(\tilde{\beta}\right)$.

## 4.2 Computation

We now provide more details on how we practically implement our estimator.

### First-Stage Nonparametric Regression

For the first-stage nonparametric estimation of $\gamma$, we adopt a machine learning estimator based on single-layer artificial neural networks, which has been widely adopted in many disciplines due to its theoretical and numerical advantages in estimating nonlinear and high dimensional functions. Clearly, model (1) naturally induces nonlinearity through the complex inequalities inside the multinomial choice model (1) with unknown forms of utility functions. Also, given that the estimation of $\gamma_{j,t,s}$ includes (time-varying) all observable product characteristics from two periods, the potentially high dimensionality of covariates also makes machine learning algorithm a suitable choice. For single-layer neural network estimators, Chen and White (1999) provides theoretical results on the convergence rates, establishing that $c_N = \left(\frac{\log N}{N}\right)^{\frac{1+2/(d+1)}{4(1+1/(d+1))}}$. On the computational side, there are also many readily usable computational packages to implement neural-network estimators. For example, in our simulation study and empirical illustration, we use the `R` package "`mlr`" by Bischl et al. (2016), which provides a front end for cross validation and hyperparameter tuning.

### Choice of the Smoothing Function $G$

Besides the requirement of Lipschitz continuity in Assumption 5, in practice we take $G$ to be bounded from above by setting $G(z) = 2\Phi\left([z]_+\right) - 1$, where $\Phi$ is the standard normal CDF. We now motivate our choice of $G$.

Recall that our identification strategy is based on the logical implication of the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$, so for identification purposes we are only interested in $\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$, i.e., whether the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$ occurs, but not in the exact magnitude of $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. However, in finite-sample, when $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is close to zero, the estimator $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is relatively more likely to have the wrong sign, so that the plug-in estimator $\mathbb{1}\left\{\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$ may induce an error of the size 1. Hence the smoothing by $G(\cdot)$ helps down-weight the observations when $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is close to zero and shrinks the magnitude of possible errors.

On the other hand, when $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is positive and large so that $\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$ can be estimated well, we do not care much about the magnitude of $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, which does not provide additional identifying information per se. By setting $G$ to be bounded from above, we dampen the effects of large $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ at the same time, so that the numerical maximization of $\hat{Q}$ is not too sensitive to

potential large but redundant variations in $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$.

**Angle-Space Reparameterization of $\mathbb{S}^{D-1}$**

In the second stage optimization of $\hat{Q}(\beta)$ over $\beta \in \mathbb{S}^{D-1}$, we work with a reparameterization of $\mathbb{S}^{D-1}$ with $(D-1)$ angles in spherical coordinates[9]. Specifically, define the angle space $\Theta$ by

$$\Theta := [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}, \tag{16}$$

and the transformation $\theta \longmapsto \beta(\theta)$ by

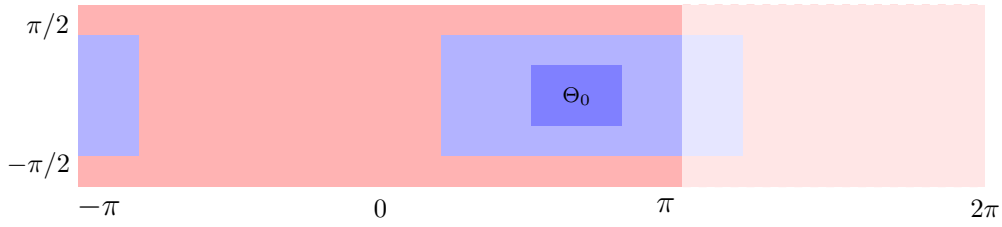
$$\beta(\theta) = \begin{cases} \beta_1(\theta) & := \cos\theta_{D-1} \ldots \cos\theta_2 \cos\theta_1, \\ \beta_2(\theta) & := \cos\theta_{D-1} \ldots \cos\theta_2 \sin\theta_1, \\ \vdots & \vdots \\ \beta_{D-1}(\theta) & := \cos\theta_{D-1} \sin\theta_{D-2}, \\ \beta_D(\theta) & := \sin\theta_{D-1}, \end{cases}$$

we now instead solves the optimization of $\hat{Q}(\beta(\theta))$ over $\Theta$, which we further equip with its natural geodesic metric $\rho_\Theta\left(\theta, \tilde{\theta}\right) := \arccos\left(\beta(\theta)' \beta\left(\tilde{\theta}\right)\right)$, which is strongly equivalent to the (imported) Euclidean distance $\left\|\beta(\theta) - \beta\left(\tilde{\theta}\right)\right\|$.

This reparameterization $(\Theta, \rho_\Theta)$ enables us to exploit the compactness and convexity of the parameter space $\Theta = [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}$, which takes the form of a hyper-rectangle. First, $(\Theta, \rho_\Theta)$ preserves all topological structure of the unit sphere, and particularly inherits the compactness of $\left(\mathbb{S}^{D-1}, \|\cdot\|\right)$, automatically satisfying the compactness condition usually imposed for extremum estimation and making it numerically feasible to initiate a grid on the whole parameter space. Second, while the unit sphere $\mathbb{S}^{D-1}$ is not convex, the new parameter space $\Theta$ becomes convex algebraically, making it computationally easy to define bisection points in the parameter space. Third, it also preserves the geometric structures of the sphere, including for instance the obvious observation that $-\pi$ and $\pi$ in the first coordinate of $\Theta$ should be treated as exactly the same point, or more rigorously, $\rho_\Theta\left((\pi - \epsilon, \theta_2, ..., \theta_{D-1}), (-\pi, \theta_2, ..., \theta_{D-1})\right) \to 0$ as $\epsilon \to 0$. This seemingly trivial prop-

[9]The idea and the motivations for using the angle-space reparameterization were also found in Manski and Thompson (1986), who however used only one angle parameter, given two pre-chosen orthogonal unit vectors on $\mathbb{S}^{D-1}$.

Figure 1: An Adaptive-Grid Algorithm




erty is nevertheless important in defining and interpreting whether certain parameter estimates converge asymptotically or not, and provides conceptual foundations for subsequent asymptotic theories.

### An Adaptive-Grid Algorithm

With the angle reparameterization, we seek to numerically compute a conservative rectangular enclosure of $\arg\min \hat{Q}(\theta)$, deploying a bisection-style grid-search algorithm that recursively shrinks and refines an *adaptive grid* to any pre-chosen precision (as defined by $\rho_\Theta$). Unlike gradient-based local optimization algorithms, our adaptive grid algorithm handles well the built-in discreteness in our sample criterion function, which has zero derivative almost everywhere, while maintains global initial coverage over the whole parameter space. While a brute-force global search algorithm is the safest choice if the dimension of product characteristics $D$ is relatively small, our adaptive-grid algorithm performs significantly faster. The essential structure of our algorithm is laid out as follows, with a corresponding illustration in Figure 1.

Step 1: Initialize a global grid $\Theta^{(1)}$ of some chosen size $M_0^{D-1}$ on $\Theta$.

Step 2: Compute $\hat{Q}(\theta)$ for each $\theta \in \Theta^{(1)}$, and select all points in $\Theta^{(1)}$ with a criterion value below the $\alpha$th-quantile in $\hat{Q}\left(\Theta^{(1)}\right) := \left\{\hat{Q}(\theta) : \theta \in \Theta^{(1)}\right\}$ into

$$\underline{\Theta}^{(1)} := \left\{\theta \in \Theta^{(1)} : \ \hat{Q}(\theta) \leq \text{quantile}_\alpha\left(\hat{Q}\left(\Theta^{(1)}\right)\right)\right\}.$$

Step 3: Take the enclosing rectangle of $\underline{\Theta}^{(1)}$, by defining $\underline{\theta}_d^{(1)} := \min^* \underline{\Theta}_d^{(1)}$ and $\overline{\theta}_d^{(1)} := \max^* \underline{\Theta}_d^{(1)}$, where $\underline{\Theta}_d^{(1)} := \left\{\theta_d : \theta \in \underline{\Theta}^{(1)}\right\}$ for each $d = 1, ..., D-1$ and the operator $\min^*$ and $\max^*$ have standard definitions of min and max except for the first dimension $d = 1$. For the first dimension, it is necessary to account for the

underlying spherical geometry and the periodicity of angles, i.e. $\theta_1 + 2\pi \equiv \theta_1$ and in particular $-\pi \equiv \pi$. This, however, is largely a programming nuisance: whenever $\underline{\Theta}_1^{(1)} \subsetneq \Theta_1^{(1)}$ crosses over at $-\pi$ and $\pi$, we can add $2\pi$ to every $\theta_1 \in \underline{\Theta}_1^{(1)}$ and obtain lower and upper bounds of $\underline{\Theta}_1^{(1)} + 2\pi$, as illustrated in Figure 1.

Step 4: We initialize a refined grid $\Theta^{(2)}$ on $\overline{\underline{\Theta}}^{(1)} := \times_{d=1}^{D-1} \left[\underline{\theta}_d^{(1)}, \overline{\theta}_d^{(1)}\right]$ of size $M_0^{D-1}$.

Step 5: Reiterate until refinement stops (falls below a certain numerical precision).

Note that the above is simply a sketch of our algorithm.[10] To be conservative, we add in buffers at each step of refinement, keep track of both outer and inner boundaries of the lower-quantile set $\underline{\Theta}^{(m)}$, and make sure that the minimizers of the criterion functions at all computed points are indeed enclosed by the set returned in the end. We find the current algorithm to be conservative and perform reasonably well in our simulation study and empirical illustration.

# 5 Simulation

In this section, we examine the finite-sample performance of our estimation method via a Monte Carlo simulation study. We start by studying the performance of the first-stage nonparametric estimator $\hat{\gamma}$ or $G(\hat{\gamma})$. Then, we show how the two-stage estimator $\hat{\beta}$ performs under various configurations of the data generating process (DGP). Finally, we investigate how our estimator performs without point identification.

### Setup of Simulation Study

For each DGP configuration, we run $M = 100$ simulations of model (1) with the following utility specification for each agent-product-time tuple $ijt$:

$$u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) = A_{i0}\left(X_{ijt}'\beta_0 + A_{ij}\right) + \epsilon_{ijt},$$

where $A_{i0}$ is an unobserved scale fixed effect that captures agent-level heteroskedasticity in utilities, and $A_{ij}$ is an unobserved location shifter specific to each agent-product

[10] Our algorithm relies heavily on the compactness and convexity of the angle space $\Theta$. Compactness allows us to start with a global grid over the whole parameter space for initial evaluations of the sample criterion function. At each step of recursion, the convexity of $\Theta$ enables us to conveniently refine the grid by separately cutting each coordinate of $\overline{\underline{\Theta}}^{(m)}$ into smaller pieces through simple division.

Table 1: Performance of First Stage Estimator $G(\hat{\gamma})$

| | $\mathbb{1}\{\hat{\gamma} > 0\}$ | $[\hat{\gamma}]_+$ | $2\Phi\left([\hat{\gamma}]_+\right) - 1$ |
|---|---|---|---|
| mean MSE | 0.1290 | 0.0221 | 0.0109 |
| max MSE | 0.1578 | 0.0254 | 0.0124 |

pair. The ability to deal with nonlinear dependence caused by the unobservable fixed effects $A$ in a robust way differentiates our method from others. To allow for such dependence, we generate correlation between the observable characteristics $\mathbf{X}_i$ and the fixed effects $\mathbf{A}_i$ via a latent variable $Z$[11]. Furthermore, we set $\beta_0 = (2, 1, ..., 1)' \in \mathbb{R}^D$ and draw $\epsilon_{ijt} \sim TIEV(0, 1)$. To summarize, for each of the $M = 100$ simulations we first generate $(\beta_0, \mathbf{X}_{it}, \mathbf{A}_i, \boldsymbol{\epsilon}_{it})$ for all $it$ combinations. Then we calculate the binary individual choice $\mathbf{Y}$ matrix according to model (1). Lastly, we compute $\hat{\beta}$ from the simulated observable data of $(\mathbf{X}, \mathbf{Y})$, and finally compare our estimator $\hat{\beta}$ with the true parameter value $\beta_0$ normalized to $\mathbb{S}^{D-1}$.

## 5.1 First-Stage Performance

We examine the performance of our first stage estimator $\hat{\gamma}$ or $G(\hat{\gamma})$. First, we calculate the true $\gamma$ or $G(\gamma)$ using the knowledge of DGP which serves as the benchmark for comparison later on. Next, we estimate $\gamma$ with only the observable data $(\mathbf{X}, \mathbf{Y})$ using single-layered neural networks and calculate the plugged-in functional $G\left(\hat{\gamma}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right)$ at each realized $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. Finally, we evaluate the performance of our estimated $G(\hat{\gamma})$ by comparing it against the true $G(\gamma)$.

We report in Table 1 both the means and the maximums of the mean squared errors (MSE) across $M$ simulations to evaluate the performance of our first stage estimator $G(\hat{\gamma})$. The header of Table 1 lists the three choices of the one-sided sign preserving function $G$. The first row, "mean MSE", reports the average MSE of $G(\hat{\gamma})$ against the true $G(\gamma)$, i.e. $\frac{1}{M}\sum_{m=1}^{M} \text{MSE}^{(m)}$ where $\text{MSE}^{(m)}$ is the MSE of $G(\hat{\gamma})$ in the $m^{th}$ simulation. The second row reports the maximum MSE of $G(\hat{\gamma})$.

From Table 1, we see that the adjusted normal CDF $2\Phi\left([\hat{\gamma}]_+\right) - 1$ performs the

[11] We draw $Z_i \sim \mathcal{N}(0, 1)$ and let $A_{i2} = [Z_i]_+$. Then, we construct $X_{ijt}^{(2)} = W_{ijt} + Z_i$ with $W_{ijt} \sim \mathcal{N}(0, 2J)$. The DGP for the rest of $\mathbf{A}$ and $\mathbf{X}$ are: $A_{i0} \sim \mathcal{U}[2, 2.5]$, $A_{i1} \equiv 0$, $A_{ij} \sim \mathcal{U}[-0.25, 0.25]$ for $j \geq 3$, $X_{ijt}^{(1)} \sim \mathcal{U}[-1, 1]$, $X_{ijt}^{(d)} \sim \mathcal{N}(0, 1)$ for $d \geq 3$.

best in terms of both mean MSE and max MSE, while the indicator function gives the worst results and that the performance of the positive part function lies somewhere in between. This is expected because when the true $\gamma$ is close to zero, it is more likely to have the estimated sign of $\hat{\gamma}$ to be different from $\gamma$. The discontinuity of the indicator function $\mathbb{1}\{\hat{\gamma} > 0\}$ at 0 magnifies this uncertainty around zero and leads to a higher MSE. When the true $\gamma$ is positive and large, it actually does not matter for our method whether the exact value of $\gamma$ is estimated well by $\hat{\gamma}$. All we need is the sign of $\hat{\gamma}$ coincides with the sign of $\gamma$ so as to obtain identifying restrictions on $\beta_0$. The adjusted normal CDF $2\Phi\left([\hat{\gamma}]_+\right) - 1$ performs the best, because it not only dampens the uncertainty in the estimated sign of $\hat{\gamma}$ near zero, but also attenuates the sensitivity to the exact value of $\hat{\gamma}_+$ relative to $\gamma_+$ when $\gamma$ is positive and large. For this reason, we will use the adjusted normal CDF function in our second stage.

## 5.2 Two-Stage Performance

We present the performance of our second stage estimator $\hat{\beta}$. First, we show the simulation results under the baseline DGP configuration, where $\beta_0$ is point-identified. Next, we study the performance of our algorithm under different numbers of individuals $N$.[12] Finally, we inspect how our estimator performs without point identification.

### Baseline Results

For the baseline configuration we set $N = 10,000, D = 3, J = 3, T = 2$. Since the sufficient conditions for point identification are satisfied under the baseline configuration, any point from the argmin set $\hat{B} := \arg\min_{\beta \in \mathbb{S}^{D-1}} \hat{Q}(\beta)$, is a consistent estimator of $\beta_0$. Specifically, we define

$$\hat{\beta}_d^u := \max \hat{B}_d, \quad \hat{\beta}_d^l := \min \hat{B}_d, \text{and} \quad \hat{\beta}_d^m := \frac{1}{2}\left(\hat{\beta}_d^u + \hat{\beta}_d^l\right)$$

for each dimension of product characteristics $d = 1, ..., D$, where $\hat{\beta}_d^u$ is the maximum value along dimension $d$ of the argmin set $\hat{B}$, $\hat{\beta}_d^l$ is the minimum value along dimension $d$ of $\hat{B}$, and $\hat{\beta}_d^m$ is the middle point along dimension $d$ of $\hat{B}$.

Table 2 summarizes the main results for the simulations under our baseline config-

[12] We also vary dimensions of observable characteristics $D$, numbers of products available $J$, and numbers of time periods $T$ and present the results in Appendix D.

Table 2: Baseline Performance

| | | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ |
|---|---|---|---|---|
| bias | $\frac{1}{M}\sum_m \left(\hat{\beta}_d^m - \beta_{0,d}\right)$ | -0.0050 | 0.0021 | 0.0006 |
| upper bias | $\frac{1}{M}\sum_m \left(\hat{\beta}_d^u - \beta_{0,d}\right)$ | 0.0015 | 0.0084 | 0.0108 |
| lower bias | $\frac{1}{M}\sum_m \left(\hat{\beta}_d^l - \beta_{0,d}\right)$ | -0.0115 | -0.0042 | -0.0096 |
| mean(u−l) | $\frac{1}{M}\sum_m \left(\hat{\beta}_d^u - \hat{\beta}_d^l\right)$ | 0.0130 | 0.0126 | 0.0205 |
| root MSE | $\left(\frac{1}{M}\sum_m \left\|\hat{\beta}^m - \beta_0\right\|^2\right)^{1/2}$ | | 0.0745 | |
| mean norm deviations | $\frac{1}{M}\sum_m \left\|\hat{\beta}^m - \beta_0\right\|$ | | 0.0648 | |

uration. In the first row of Table 2 we use the middle value $\hat{\beta}^m$ along each dimension of set estimator $\hat{B}$ to calculate the average bias against the true $\beta_0$ across all $M = 100$ simulations. The bias is very small across all three dimensions with a magnitude between -0.0050 and 0.0021. The next two rows show the biases in estimating $\beta_{0,d}$ using $\hat{\beta}_d^u$ and $\hat{\beta}_d^l$ respectively and the biases are again close to zero. The fourth row of Table 2 measures the average width of the set estimator $\hat{B}$ along each dimension. It is relatively tight compared to the magnitude of $\beta_0$. In the second part of Table 2 we report the root MSE (rMSE) and mean norm deviations (MND) using $\hat{\beta}^m$. Our proposed algorithm is able to achieve a low rMSE and MND.

**Results Varying $N$**

We vary $N$ while maintaining $D = 3, J = 3, T = 2$ to show how our method performs under different sample sizes. In addition to our baseline setup with $N = 10,000$, we calculate mean absolute deviation (MAD), average size of the estimated set, rMSE and MND for $N = 4,000$ and $N = 1,000$. Results are summarized in Table 3.

From Table 3, it is clear that a larger $N$ helps with overall performance. MAD decreases from 0.0694 to 0.0077 when $N$ increases from $1,000$ to $10,000$. The average size of the estimated sets, the rMSE, and the MND show a similar pattern. However, even with a relatively small $N = 1,000$ the result from our method is still quite informative and accurate, with the average size of the estimated set and the MND being equal to 0.1076 and 0.1405, respectively. We emphasize that here the total number of time periods $T$ is set to a minimum of 2. Our method can extract information from

Table 3: Performance under Varying $N$

| | $\sum_d \lvert\text{bias}_d\rvert$ | $\sum_d \text{mean(u-l)}_d$ | rMSE | MND |
|---|---|---|---|---|
| $N = 10,000$ | 0.0077 | 0.0461 | 0.0745 | 0.0648 |
| $N = 4,000$ | 0.0174 | 0.0715 | 0.1006 | 0.0884 |
| $N = 1,000$ | 0.0694 | 0.1076 | 0.1690 | 0.1405 |
| | $\left(\frac{N}{1,000}\right)^{1/2}$ | $\left(\frac{N}{1,000}\right)^{1/3}$ | $\frac{\text{rMSE}_{1000}}{\text{rMSE}_N}$ | $\frac{\text{MND}_{1000}}{\text{MND}_N}$ |
| $N = 10,000$ | 3.16 | 2.15 | $\frac{0.1690}{0.0745} \approx 2.27$ | $\frac{0.1405}{0.0648} \approx 2.17$ |
| $N = 4,000$ | 2.00 | 1.59 | $\frac{0.1690}{0.1006} \approx 1.68$ | $\frac{0.1405}{0.0884} \approx 1.59$ |

each of the $T(T-1)$ ordered pairs of time periods, which increase quadratically with $T$. See Appendix D for results with larger $T$.

Next, we numerically investigate the speed of convergence of our method when we increase sample size $N$ from $1,000$ to $4,000$ and $10,000$ in the second part of Table (3). Compared with the case of $N_0 = 1,000$, the relative ratios of rMSE are 1.68 for $N = 4,000$ and 2.27 for $N = 10,000$, both of which lie between $(N/N_0)^{1/3}$ and $(N/N_0)^{1/2}$. A similar pattern is also found for calculations based on MND. These results indicate that our method achieves a convergence rate slower than the $N^{-1/2}$ but slightly faster than the $N^{-1/3}$ rate.

**Estimation without Point Identification**

We now investigate the performance of our estimator under specifications where point identification fails. To make things comparable, we fix $(N, D, J, T)$ as in the baseline case, but we modify the configuration in two different ways. We maintain the point identification of $\beta_0$ in one setting but lose the point identification in the other[13]. We deliberately control the location and scale of each variable to be comparable across the two configurations, with the only differences being the presence of discreteness and boundedness of supports. When point identification fails, we compute the set

[13]Specifically, we set $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $X_{ijt}^{(1)} \sim \mathcal{U}[-1,1]$, $X_{ijt}^{(2)} = Z_i + \mathcal{N}(0,6)$, and $X_{ijt}^{(3)} \sim \mathcal{N}(0,1)$ for the point identified case. For the DGP without point identification, we let $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $X_{ijt}^{(1)} \sim \mathcal{U}\{-1,1\}$, $X_{ijt}^{(2)} = Z_i + \mathcal{U}\left(-\sqrt{6}, \sqrt{6}\right)$, and $X_{ijt}^{(3)} \sim \mathcal{U}[-1,1]$. The key difference is that $X$ in the latter DGP has a bounded support.

Table 4: Performance with and without Point ID: Further Examination

| point ID ? | $\hat{c}$ | rMSE | | | MND | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}^m$ | $\hat{\beta}^u$ | $\hat{\beta}^l$ | $\hat{\beta}^m$ | $\hat{\beta}^u$ | $\hat{\beta}^l$ |
| **(i)** yes | - | 0.0770 | 0.0789 | 0.0795 | 0.0661 | 0.0685 | 0.0697 |
| **(ii)** no | 0.01 | 0.0872 | 0.0880 | 0.0894 | 0.0753 | 0.0767 | 0.0775 |
| | 0.1 | 0.0860 | 0.0929 | 0.0939 | 0.0737 | 0.0833 | 0.0832 |
| | 1 | 0.0790 | 0.1268 | 0.1447 | 0.0668 | 0.1207 | 0.1295 |

estimator $\hat{B}_{\hat{c}}$ of (15) with $\hat{c} > 0$. Table 4 contains simulation results under the two configurations, with different choices of $\hat{c}$ when point identification fails. [14]

In Table 4 , we calculate the rMSE and MND of the upper bound $\hat{\beta}^u$, the lower bound $\hat{\beta}^l$ and the middle point $\hat{\beta}^m$ of the (approximate) argmin sets$\hat{B}_{\hat{c}}$ (with $\hat{c} = 0$ under point identification and three choices of $\hat{c}$ under partial identification) with respect to the true normalized parameter $\beta_0$. Across rows in (i) and (ii), we see that the lack of point identification does negatively affect the performance of our estimates, but the impact is limited to a moderate degree. Within rows in (ii), we observe that, as expected, a more conservative choice of the constant $\hat{c}$ worsens performances of the upper and lower bounds by enlarging the estimated sets; in the meanwhile, it appears that the size (and the performance) of our estimator based on $\hat{\beta}^m$ is not terribly sensitive to the choice of $\hat{c}$.

# 6 Empirical Illustration

## 6.1 Data and Methodology

As an empirical illustration, we apply our method to the Nielsen Retail Scanner Data on popcorn sales to explore the effects of display promotion effects. The Nielsen Retail Scanner Data contains weekly information on store-level price, sales and display promotion status generated by about 35,000 participating retail store with point-of-sale systems across the United States. Among a huge variety of products covered by

[14] Specifically, noting that $c_N \log N \leq N^{-1/4} \log N \approx 0.92 \leq 1$ for $N = 10,000$, we set $\hat{c} = 0.01$, 0.1 and 1, respectively.

Table 5: Empirical Application: Summary Statistics

| | mean | s.d. | min | max |
|---|---|---|---|---|
| DMA-level Market Share $s_{ijt}$ | 25.00% | 21.59% | 0.07% | 96.69% |
| $\text{Price}_{ijt}$ | 0.4924 | 0.1803 | 0.1094 | 1.3587 |
| $\text{Promo}_{ijt}$ | 0.0282 | 0.0377 | 0.0000 | 0.5000 |
| $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ | 0.0136 | 0.0203 | 0.0000 | 0.4505 |

the Nielsen data, we choose to focus on popcorn for two reasons. First, purchases of popcorn are more likely to be driven by temporary urges of consumption without too much dynamic planning. Second, there is good variation in the display promotion status of popcorn, which enables us to estimate how important special in-store displays affect consumer's purchase decisions.

We aggregate the store level data to the $N = 205$ designated market area (DMA) level for year 2015. We focus on the top 3 brands ranked by market share, aggregate the rest into a fourth product "all other products", and allow an outside option of "no purchase". We calculate the dependent variable "market share" for each of the $J = 5$ brands. The observed product characteristics $\mathbf{X}$ include price, promotion status and their interaction term[15]. The summary statistics of the variables discussed above are provided in Table 5.

To describe the methodology, we use the observed DMA-level market shares as an estimate of $s_{ijt} = \mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right]$. Under the strong stationarity assumption, we run the first-stage estimation of

$$\mathbb{E}\left[s_{ijt} - s_{ijs} \middle| \mathbf{X}_{i,ts}\right] = \int \left(\mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right] - \mathbb{E}\left[y_{ijs} \middle| \mathbf{X}_{is}, \mathbf{A}_i\right]\right) d\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{i,ts}\right).$$

Specifically, we nonparametrically regress $(s_{ijt} - s_{ijs})$ on $\mathbf{X}_{i,ts}$ using single-layered neural networks from the `mlr` package in `R`, and obtain an estimator $\hat{\gamma}_j$ of $\gamma_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[s_{ijt} - s_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right]$. Then, we plug $\hat{\gamma}$ into our second-stage algorithm and

[15] We calculate $\text{Price}_{ijt}$ as the weighted average unit price of all $UPC$s of the brand $j$ in DMA $i$ during week $t$. In the Nielsen data we find two variables related to promotion: display and feature. Due to their similarity, we calculate $\text{Promo}_{ijt}$ as $(\text{feature} \vee \text{display})_{ijt}$. The interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ is included in $X$ to show the effect of promotion on the price elasticity of consumers.

Table 6: Empirical Application: Estimation Results

| | $\hat{\beta}^m_{\hat{c}=0}$ | $\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}=0}$ | $\hat{\beta}^m_{\hat{c}=0.014}$ | $\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}=0.014}$ |
|---|---|---|---|---|
| $\text{Price}_{ijt}$ | -0.9681 | [-0.9687, -0.9677] | −0.9236 | [-0.9711, -0.8761] |
| $\text{Promo}_{ijt}$ | 0.1970 | [ 0.1861, 0.2078] | 0.1565 | [ 0.0662, 0.2469] |
| $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ | 0.1550 | [ 0.1399, 0.1700] | 0.2731 | [ 0.0687, 0.4776] |

Table 7: Empirical Illustration: Comparison of Results

| | $\hat{\beta}^m$ | $\hat{\beta}^{CyclicMono}$ | $\hat{\beta}^{OLS}$ | $\hat{\beta}^{OLS-FE}$ | $\hat{\beta}^{MLogit-FE}$ |
|---|---|---|---|---|---|
| $\text{Price}_{ijt}$ | -0.9236 | -0.3781 | 0.0240 | -0.3803 | -0.8511 |
| $\text{Promo}_{ijt}$ | 0.1565 | -0.0567 | 0.5760 | 0.5978 | 0.4589 |
| $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ | 0.2731 | 0.9240 | -0.8171 | -0.7057 | -0.2552 |

compute the (approximate) argmin set $\hat{B}_{\hat{c}}$.

## 6.2 Results and Discussion

We report our estimation results in Table 6. $\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}}$ corresponds to the lower and upper bounds of the (approximate) argmin set $\hat{B}_{\hat{c}}$, while $\hat{\beta}^m_{\hat{c}} := \frac{1}{2}\left(\hat{\beta}^l_{\hat{c}} + \hat{\beta}^u_{\hat{c}}\right)$ corresponds to the middle point. We show both the exact argmin set ($\hat{c} = 0$) and the approximate argmin set with $\hat{c} = 0.01 \times N^{-\frac{1}{4}} \log(N) \approx 0.014$ for $N = 205$. The estimated coefficients for Price (negative) and Promo (positive) are clearly consistent with economic intuitions.

The most interesting result is the positive estimated coefficient on the interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$. An intuitive explanation for the positive sign is that by displaying certain products in front rows, consumers no longer see the price tags of these products adjacent to those of their competitors, and consequently become less price-sensitive for these specially promoted products.

To further illustrate the advantages of our method, we compare our $\hat{\beta}^m$ with the estimates obtained through four other different popular methods, i.e. Cyclic Mono-

tonicity (CM) based on Shi, Shum, and Song (2018)[16], classic OLS, OLS with scalar-valued fixed effects (OLS-FE) and the multinomial logit with fixed effects (MLogit-FE). Results (normalized to $\mathbb{S}^{D-1}$) are summarized in Table 7.

The OLS regression result shows that the estimated coefficient on $\text{Price}_{ijt}$ is 0.0240, which is counterintuitive and unreasonable. Moreover, as explained before, displaying the product at the front row of the store will likely make consumers less price sensitive, implying a positive coefficient for $\text{Price}_{ijt}\times\text{Promo}_{ijt}$. However, the estimated coefficients for the interaction term using OLS, OLS-FE and MLogit-FE are all negative, contrary to that intuition. Finally, the CM-based method reports a small but negative coefficient of -0.0567 for $\text{Promo}_{ijt}$, which could be hard to rationalize.

We regard the contrast between our result and the results obtained in these alternative methods as an empirical illustration that by accommodating more flexible forms of unobserved heterogeneity, through the arbitrary dimensional fixed effects that are allowed to enter into consumers' utility functions in an additively nonseparable way, our method is able to produce economically more reasonable results.

## 6.3 A Possible Explanation via Monte-Carlo Simulations

In this section, we propose a possible explanation to the empirical findings in Table 7 via a Monte Carlo simulation. Recall that "Promo" captures whether a product gains increased exposure by being highlighted by stores. We argue that the negative estimated coefficients obtained in traditional methods in Table 7 for $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ may be caused by a positive correlation between display promotion and unobserved index sensitivity, the latter of which enters the utility function nonlinearly.

Specifically, suppose the utility function can be written as

$$u_{ijt} = A_{ij} \times \left(X_{ijt}^{'}\beta_0\right) + \epsilon_{ijt}, \tag{17}$$

where $X_{ijt}$ contains Price, Promo, and Price×Promo, $A_{ij}$ is the $ij-$specific fixed effect which may capture index sensitivity (which can be thought as inversely related to unobserved brand loyalty), and $\epsilon_{ijt}$ is the exogenous random shock. Suppose $A_{ij}$ and $\text{Promo}_{ijt}$ is positively correlated, which is reasonable because marketing managers with their expertise are more likely to promote products to which consumers are more price and promotion sensitive. Thus, traditional estimation methods that base

[16] We used 2-week cycles for all available weeks in the data for the CM method.

Table 8: Percentage of Correct Signs of Estimated Coefficients

| $\alpha$ | $\hat{\beta}^{m}$ | $\hat{\beta}^{CyclicMono}$ | $\hat{\beta}^{OLS}$ | $\hat{\beta}^{OLS-FE}$ | $\hat{\beta}^{MLogit-FE}$ |
|---|---|---|---|---|---|
| 0.15 | 96% | 0% | 0% | 0% | 6% |
| 0.30 | 97% | 0% | 0% | 0% | 0% |
| 0.50 | 82% | 0% | 0% | 0% | 0% |

on linearity would be unable to detect such pattern and wrongly attribute the effect on price elasticities from $A_{ij}$ to Promo.

To provide some numerical evidence of the claim, we run the following Monte Carlo simulation. We let $\beta_0 = (-4, 2, 2)^{'}$, $Z \sim \mathcal{U}[0,1]$, $A_{ij} = Z + 1$, and $\epsilon_{ijt} \sim TIEV(0,1)$. For $X_{ijt}$ vector, we draw $X_{ijt}^{(1)} \sim \mathcal{U}[0,4]$ and $W \sim \mathcal{U}[0,1]$, and let $X_{ijt}^{(2)} = (1-\alpha) \times W + \alpha \times Z$ and $X_{ijt}^{(3)} = X_{ijt}^{(1)} \times X_{ijt}^{(2)}$. We emphasize that $X_{ijt}^{(2)}$(Promo) is positively correlated with $A_{ij}$ through $Z$, with $\alpha$ measuring the strength of the correlation. We consider three values of $\alpha$: $0.15, 0.3$ and $0.5$.

We run 100 simulations for each of the five methods in Table 7 to estimate $\beta_0$. To replicate the data structure of the empirical exercise, we set $N = 205$, $D = 3$, $J = 4$, and $T = 52$. We report in Table 8 the percentage of simulations that the corresponding method is able to generate correct signs for all coordinates of $X_{ijt}$.

The percentages that our proposed method is able to generate correct signs for all coordinates of $X_{ijt}$ for $\alpha = 0.15,\ 0.3$, and $0.5$ are 96%, 97%, and 82%, respectively. The accuracy of the estimator is negatively affected by the correlation between $X_{ijt}^{(2)}$ (Promo) and $A_{ij}$ (multiplicative fixed effect). None of the other methods in Table 8 generates estimates of $\beta_0$ with correct signs. It is worth mentioning that the CM-based method requires $A_{ij}$ entering the utility function linearly, which is violated in our DGP in (17). Apparently, all these other models than ours, due to their additive separable structure, completely ignore the positive dependence between the observable covariate $X_{ijt}^{(2)}$ (promotion) and the multiplicative fixed effect $A_{ij}$, thus producing biases in their estimates.

Intuitively, since products with larger $A_{ij}$ are more likely to be promoted $\left(X_{ijt}^{(2)} = 1\right)$ by the selection of marketing managers, the average effective price sensitivity of promoted products tend to be larger than those products not promoted. This drives those estimators that ignore such confounding selection effects to produce

a negative coefficient on the interaction term $X_{ijt}^{(1)} \times X_{ijt}^{(2)}$ (Price $\times$ Promo), as found in the empirical illustration (Table 7). In contrast, our method handles such *non-additive* dependence between observable characteristics and unobserved fixed effects reasonably well, illustrating the robustness of our methods.

# 7 Monotone Multi-Index Models

We now present a general framework under which our identification strategy is applicable, using the notation of Ahn, Ichimura, Powell, and Ruud (2018, AIPR thereafter):

$$\gamma\left(\mathbf{X}_i\right) = \phi\left(\mathbf{X}_i\beta_0\right) \tag{18}$$

in which: $(y_i, \mathbf{X}_i)_{i=1}^N$ constitutes a random sample of $N$ observations on a scalar[17] random variable $y_i$ and a $J \times D$ random matrix $\mathbf{X}_i$. $\gamma\left(\overline{\mathbf{X}}\right) = \mathcal{T}\left(F_{y_i|\mathbf{X}_i=\overline{\mathbf{X}}}\left(\cdot\right)\right)$ is a real variable defined as a known functional $\mathcal{T}$ of the conditional distribution of $y_i$ given $\mathbf{X}_i = \overline{\mathbf{X}}$. A leading example is to set $\gamma\left(\mathbf{X}_i\right) := \mathbb{E}\left[y_i|\,\mathbf{X}_i\right]$, so that model (18) becomes a conditional moment condition; however, this is not necessary. $\phi : \mathbb{R}^J \to \mathbb{R}$ is an unknown real-valued function. $\beta_0 \in \mathbb{R}^D \backslash \{\mathbf{0}\}$ is the unknown finite-dimensional parameter of interest. Again, we normalize $\beta_0 \in \mathbb{S}^{D-1}$, as $\beta_0$ is at best identified up to scale given that $\phi$ is an unknown function. As in Lee (1995), Powell and Ruud (2008) and AIPR, model (18) restricts the dependence of $\gamma\left(\mathbf{X}_i\right)$ on the matrix $\mathbf{X}_i$ to the $J$ linear parametric indexes $\mathbf{X}_i\beta_0 \equiv \left(X_{ij}'\beta_0\right)_{j=1}^J$.[18]

A noteworthy difference of model (18) from the setup in AIPR is that we take $\gamma\left(\mathbf{X}_i\right)$ here to be scalar-valued, while AIPR require their $\gamma\left(\mathbf{X}_i\right)$ to have dimension, using their notation $R$, no smaller than $J$. This "order condition" $R \geq J$ is necessary for their vector-valued function $\phi$ to admit a left-inverse $\phi^{-1}$ such that $\phi^{-1}\left(\gamma\left(\mathbf{X}_i\right)\right) = \mathbf{X}_i\beta_0$, which constitutes the foundation for their subsequent analysis. In contrast, we impose no such order condition for the sake of invertibility, as we will not rely on

[17] Similar to AIPR, the dimension of $y_i$ is largely irrelevant to the analysis of model (18): it is the dimension of $\gamma$ that matters. Nevertheless, for the clarity of presentation, we take $y_i$ to be a scalar.

[18] Note that model (18) is WLOG relative to the following seemingly more general formulation, in which $\beta_0$ is explicitly allowed to be heterogeneous across the $J$ rows of $\mathbf{X}_i$: $\gamma\left(\mathbf{X}_i\right) = \phi\left(\left(X_{ij}'\beta_{0j}\right)_{j=1}^J\right)$, where $\beta_0 := \left(\beta_{01}', ..., \beta_{0J}'\right)'$ is a $\sum_{j=1}^J D_j$-dimensional vector. This, however, could be readily incorporated in model (18) by appropriately redefining $\tilde{\mathbf{X}}_i$ to obtain the representation $\gamma\left(\tilde{\mathbf{X}}_i\right) = \phi\left(\tilde{\mathbf{X}}_i\beta_0\right)$ as in model (18).

invertibility at all. Instead, we impose the following monotonicity assumption.

**Assumption 7** (Weak Monotonicity)**.** *$\phi$ is nondegenerate and nondecreasing in each of its $J$ arguments on Supp $(\mathbf{X}_i\beta_0) \subseteq \mathbb{R}^J$.*

With no other restrictions besides Assumption 7 on the unknown function $\phi$, model (18) builds in the fundamental lack of additive separability across the parametric indexes. As demonstrated in Section 2, the key idea developed below for the general multi-index model (18) naturally applies to the analysis of the panel multinomial choice model under complete lack of additive separability.

We now provide a few illustrative examples for model (18) that satisfy Assumption 7 beyond multinomial choice settings.

**Example 1** (Sample Selection Model)**.** Consider the sample selection model studied by Heckman (1979), where $y_i = y_i^* \cdot d_i$ with $y_i^* = W_i^{'}\mu_0 + u_i$ and $d_i = \mathbb{1}\left\{Z_i^{'}\lambda_0 + v_i \geq 0\right\}$. We observe $(y_i, W_i, Z_i)$ but not $y_i^*$. Suppose $(u_i, v_i) \perp (X_i, Z_i)$ and the joint distribution of $(u_i, v_i)$ is bivariate normal with a positive correlation. Then we have

$$\mathbb{E}\left[y_i \middle| W_i, d_i = 1\right] = W_i^{'}\mu_0 + \mathbb{E}\left[u_i \middle| v_i \geq -Z_i^{'}\lambda_0\right] =: \phi\left(W_i^{'}\mu_0, -Z_i^{'}\lambda_0\right).$$

By taking $X_i := (W_i, Z_i, d_i)$ and $\beta_0 := (\mu_0, \lambda_0)$, we may easily rewrite the model in the formulation of model (18) with Assumption 7 satisfied.

**Example 2** (Dyadic Network Formation Model under Nontransferable Utilities)**.** Consider the following simple dyadic network formation model under nontransferable utilities (NTU):

$$D_{ij} = \mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ij}^{'}\gamma_0 \geq \epsilon_{ij}\right\} \mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ji}^{'}\gamma_0 \geq \epsilon_{ji}\right\}, \tag{19}$$

where $W_{ij} \equiv W_{ji}$ denotes some symmetric observable characteristics between a pair of individuals $ij$, $(Z_{ij}, Z_{ji})$ represent some asymmetric observable characteristics between $ij$, and $(\epsilon_{ij}, \epsilon_{ji})$ denote some potentially asymmetric idiosyncratic shocks to $i$'s and $j$'s utilities from linking with each other. The observed binary variable $D_{ij} \equiv D_{ji}$ of an undirected link between $ij$ is determined jointly by two threshold-crossing conditions, interpreted as the requirement of mutual consent in the establishment of a link between $ij$. Clearly, we have

$$\mathbb{E}\left[D_{ij} \middle| W_{ij}, Z_{ij}, Z_{ji}\right] = \phi\left(W_{ij}^{'}\mu_0, Z_{ij}^{'}\gamma_0, Z_{ji}^{'}\gamma_0\right),$$

which falls under model (18) with Assumption 7 satisfied. It is worth noting that the NTU setting, which is a highly plausible feature in the formation of social networks, naturally induces lack of additive separability via the multiplication of two threshold-crossing conditions, even if we have a fully additive specification inside each threshold-crossing condition as in (19). Hence, the NTU setting provides a micro-founded motivation for confronting nonseparability, which our key method is well suited to deal with.

In a companion paper (Gao, Li, and Xu, 2020), we study a related but more complicated model of dyadic link formation with unobserved degree heterogeneity:

$$D_{ij} = \mathbb{1}\left\{u\left(W_{ij}'\beta_0, A_i, A_j\right) \geq \epsilon_{ij}\right\} \mathbb{1}\left\{u\left(W_{ij}'\beta_0, A_j, A_i\right) \geq \epsilon_{ji}\right\},$$

where $A_i$ and $A_j$ are scalar-valued individual "fixed effects" that represent each individual's unobserved heterogeneity in sociability. The involvement of the two-way fixed effects in this network formation setting adds further complications relative to the panel multinomial choice model considered in this paper, and we propose a new method, called *logical differencing*, to cancel out the two-way fixed effects, by constructing an observable event that contains the intersection of two mutually exclusive restrictions on the fixed effects. Nevertheless, the logical contraposition of multivariate monotonicity remains a convenient device for our identification arguments.

**Proposition 2** (General Identifying Restriction)**.** *Under model (18) with Assumption 7, for any* $\overline{\mathbf{X}}, \underline{\mathbf{X}} \in Supp\left(\mathbf{X}_i\right)$, $\gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right)$ *implies NOT* $\left\{\left(\overline{X}_j - \underline{X}_j\right)\beta_0 \leq 0,\ \forall j = 1, ..., J\right\}$.

Proposition 2 generalizes our key identification result (Theorem 1). Notice that Proposition 2 applies to all functionals $\gamma$ on the conditional distribution $y_i|\,\mathbf{X}_i$ that satisfy the monotonicity assumption. Besides conditional expectations, there are many models where conditional quantiles or higher-order conditional moments are more natural choices of $\gamma$. In some cases where the whole conditional distribution $y_i|\,\mathbf{X}_i$ can be ranked by first-order or second-order stochastic dominance, we may aggregate the identifying information from many choices of $\gamma$ into a joint restriction on $\beta_0$. We leave a further analysis of this topic to future research.

# 8 Conclusion

This paper proposes a simple and robust method for semiparametric identification and estimation in a panel multinomial choice model, exploiting the standard notion of multivariate monotonicity in an index vector of observable characteristics.

Our key identification strategy using logical contraposition of multivariate monotonicity is very simple, but it is exactly this conceptual simplicity that lends us the ability to accommodate infinite dimensionality of unobserved heterogeneity and lack of additive separability in consumer preferences. As the validity of this methodology essentially relies on nothing but monotonicity in a parametric index structure, it should be more widely applicable beyond the multinomial choice settings we consider.

However, a more comprehensive or in-depth investigation of whether and how this strategy can be adapted to the peculiarities of specific economic problems still requires a substantial amount of future work to be done. For applications in industrial organization, it might be worthwhile to inspect whether certain form of monotonicity can be preserved, at least approximately, in the presence of additional features, such as random coefficients and time-varying endogeneity, under certain conditions. In connection to microeconomic theory, it might also be interesting to investigate whether theoretical results on monotone comparative statics can be combined with our monotonicity-based method to provide a venue of identification and estimation in endogenous economic systems.

Furthermore, the asymptotic theory of the semiparametric estimator considered in this paper turns out to be interesting even in a binary choice model with point identification, as it features a nonstandard interplay between the nonsmooth sample criterion and the effective smoothing asymptotically provided by the first-stage estimator. Given that the asymptotic theory of such estimators is of independent interest and is better studied under different settings and notations, we refer interested readers to Gao and Xu (2020) for more details.

# Appendix A: Proof of Theorem 2

We first prove two lemmas before formally proving Theorem 2.

**Lemma 3.** *$Q : \mathbb{S}^{d-1} \to \mathbb{R}_+$ is continuous.*

*Proof.* Recalling that $\cdot v_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right)=\overline{X}_k-\underline{X}_k/\left\|\overline{X}_k-\underline{X}_k\right\|$ whenever $\overline{X}_k\neq\underline{X}_k$ while $v_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right)=0$ when $\overline{X}_k=\underline{X}_k$, we have

$$\begin{aligned}G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right)=&G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\prod_{k=1}^{J}\mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}}\left(X_{ikt}-X_{iks}\right)'\beta\geq 0\right\}\\=&G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\prod_{k=1}^{J}\mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}}v_k\left(\mathbf{X}_{it}-\mathbf{X}_{is}\right)'\beta\geq 0\right\}\end{aligned}$$

which is continuous in $\beta$ with probability one, since $v_k\left(\mathbf{X}_{it}-\mathbf{X}_{is}\right)$ has no mass point except possibly at $\mathbf{0}$, in which case the indicator degenerates to a constant over $\beta\in\mathbb{S}^{d-1}$. Since $\mathbf{X}_{i,ts}$ is i.i.d. across $i$, $\mathbb{S}^{d-1}$ is compact, and the indicator function is bounded, all conditions for Lemma 2.4 in Newey and McFadden (1994) are satisfied, by which we conclude that $Q=\sum_{j,t,s}Q_{j,t,s}$ is continuous on $\mathbb{S}^{d-1}$. □

**Lemma 4.** *Under Assumptions 2, 5 and 6,* $\sup_{\beta\in\mathbb{S}^{d-1}}\left|\hat{Q}\left(\beta\right)-Q\left(\beta\right)\right|=O_p\left(c_N\right)$.

*Proof.* We first prove the convergence of $\hat{Q}_{j,t,s}\left(\beta\right)$ to $Q_{j,t,s}\left(\beta\right)$ for each $(j,t,s)$. For each generic deterministic function $\tilde{\gamma}_{j,t,s}$, define

$$\begin{aligned}Q_{j,t,s}\left(\beta,\tilde{\gamma}\right)&:=\mathbb{E}\left[G\left(\tilde{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right)\right],\\\hat{Q}_{j,t,s}\left(\beta,\tilde{\gamma}\right)&:=\frac{1}{n}\sum_{i=1}^{n}G\left(\tilde{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right).\end{aligned}$$

so that $\hat{Q}_{j,t,s}\left(\beta\right)=\hat{Q}_{j,t,s}\left(\beta,\tilde{\gamma}_{j,t,s}\right)$ and $Q_{j,t,s}\left(\beta\right)=Q_{j,t,s}\left(\beta,\gamma\right)$. For notational simplicity we suppress the subscript $(j,t,s)$ for the moment.

Defining $\mathcal{Q}:=\left\{G\left(\tilde{\gamma}\left(\overline{\mathbf{X}}\right)\right)\lambda\left(\mathbf{X}_{i,ts};\beta\right):\tilde{\gamma}\in\Gamma,\beta\in\mathbb{S}^{d-1}\right\}$, we first argue that $\mathcal{Q}$ is a $\mathbb{P}$-Donsker class based on Van Der Vaart and Wellner (1996). First, it is easy to show by Assumption 5 that $G\left(0\right)=0$, which together with the Lipschitz continuity of $G$, we have $\mathbb{E}\left[G^2\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right]\leq M\mathbb{E}\left[\tilde{\gamma}^2\left(\mathbf{X}_i\right)\right]<\infty$ and $\mathbb{E}\left|G\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right|\leq\mathbb{E}\left|\tilde{\gamma}\left(\mathbf{X}_i\right)\right|\leq\sup_{\tilde{\gamma}\in\Gamma}\mathbb{E}\left|\tilde{\gamma}\left(\mathbf{X}_i\right)\right|<\infty$. Then, as $\Gamma$ is $\mathbb{P}$-Donsker, $G\circ\tilde{\gamma}$ must also be $\mathbb{P}$-Donsker. Second, recall that $\lambda\left(\mathbf{X}_{i,ts};\beta\right)$ is the product of indicators of half planes, while the collection of $\mathbb{1}\left\{\left(\overline{X}_k-\underline{X}_k\right)'\beta\geq 0\right\}$ over $\beta\in\mathbb{S}^{d-1}$ is a well-known VC Class of functions (sets) and is thus $\mathbb{P}$-Donsker. Finally, since the indicator function is uniformly bounded and $\sup_{\tilde{\gamma}\in\Gamma}\mathbb{E}\left|G\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right|<\infty$, we conclude that $\mathcal{Q}$ is also $\mathbb{P}$-Donsker:

$$\sup_{\beta\in\mathbb{S}^{d-1}}\sup_{\tilde{\gamma}\in\Gamma}\left|\hat{Q}\left(\beta,\tilde{\gamma}\right)-Q\left(\beta,\tilde{\gamma}\right)\right|=O_p\left(N^{-\frac{1}{2}}\right).\tag{20}$$

Next, by Assumption 4, we have

$$\begin{aligned}\sup_{\beta\in\mathbb{S}^{d-1}}\left|Q\left(\beta,\hat{\gamma}\right)-Q\left(\beta,\gamma\right)\right| \leq & \sup_{\beta\in\mathbb{S}^{d-1}}\int\left|G\left(\hat{\gamma}\left(\overline{\mathbf{X}}\right)\right)-G\left(\gamma\left(\overline{\mathbf{X}}\right)\right)\right|\lambda_j\left(\overline{\mathbf{X}};\beta\right)d\mathbb{P}\left(\overline{\mathbf{X}}\right) \\ \leq & M\sqrt{\int\left(\hat{\gamma}\left(\overline{\mathbf{X}}\right)-\gamma\left(\overline{\mathbf{X}}\right)\right)^2 d\mathbb{P}\left(\overline{\mathbf{X}}\right)}=O_p\left(c_N\right) \qquad (21)\end{aligned}$$

by Lipschitz continuity of $G$, $|\lambda_j|\leq 1$ and Cauchy-Schwarz. Combining (20) and (21), we have

$$\begin{aligned}\sup_{\beta\in\mathbb{S}^{d-1}}\left|\hat{Q}\left(\beta,\hat{\gamma}\right)-Q\left(\beta,\gamma\right)\right| \leq & \sup_{\beta\in\mathbb{S}^{d-1}}\sup_{\tilde{\gamma}\in\Gamma}\left|\hat{Q}\left(\beta,\tilde{\gamma}\right)-Q\left(\beta,\tilde{\gamma}\right)\right|+\sup_{\beta\in\mathbb{S}^{d-1}}\left|\hat{Q}\left(\beta,\hat{\gamma}\right)-Q\left(\beta,\hat{\gamma}\right)\right| \\ = & O_p\left(N^{-\frac{1}{2}}\right)+O_p\left(c_N\right)=O_p\left(c_N\right)\end{aligned}$$

since $N^{-\frac{1}{2}}=O_p\left(c_N\right)$ for nonparametric estimators. Summing over all $(j,t,s)$, we have $\sup_{\beta\in\mathbb{S}^{d-1}}\left|\hat{Q}\left(\beta\right)-Q\left(\beta\right)\right|=O_p\left(c_N\right)$. □

**Main Proof of Theorem 2**

*Proof.* We verify Condition C.1 in Chernozhukov, Hong, and Tamer (2007, CHT thereafter) so as to apply their Theorem 3.1. Condition C.1(a) on the nonemptiness and compactness of parameter space is satisfied given Theorem 1. Condition C.1(b) on the continuity of the population criterion function $Q$ is proved by Lemma 3. Condition C.1(c) on measurability of the sample criterion function is satisfied by its construction. Condition C.1(d)(e) regarding the uniform convergence of $Q_n$ are satisfied by Lemma 4. Hence Theorem 3.1.(1) in CHT implies the Hausdorff consistency of $\hat{B}$. The consistency of the point estimator under the additional assumption of point identification (i.e., $B_0$ is a singleton) follows from Theorem 3.2 of CHT. □

# References


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

Ahn, H., H. Ichimura, J. L. Powell, and P. A. Ruud (2018): "Simple Estimators for Invertible Index Models," *Journal of Business & Economic Statistics*, 36.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected Substitutes and Invertibility of Demand," *Econometrica*, 81, 2087–2111.

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

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones (2016): "mlr: Machine Learning in R," *The Journal of Machine Learning Research*, 17, 5938–5942.

Browning, M. and J. Carro (2007): *Heterogeneity and Microeconometrics Modeling*, Cambridge University Press, vol. 3 of *Econometric Society Monographs*, 47–74.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

Chen, X. and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., I. Fernández-Val, and W. K. Newey (2019): "Nonseparable Multinomial Choice Models in Cross-Section and Panel Data," *Journal of econometrics*, 211, 104–116.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Fan, J., F. Han, and H. Liu (2014): "Challenges of Big Data Analysis," *National Science Review*, 1, 293–314.

Fox, J. T. (2007): "Semiparametric Estimation of Multinomial Discrete-Choice Models Using a Subset of Choices," *The RAND Journal of Economics*, 38, 1002–1019.

Gao, W. Y., M. Li, and S. Xu (2020): "Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities," Working Paper.

Gao, W. Y. and S. Xu (2020): "Two-Stage Maximum Score Estimator in Monotone Index Models," Working Paper.

Han, A. K. (1987): "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (2001): "Micro data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Honoré, B. E. and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for The Binary Response Model," *Econometrica: journal of the Econometric Society*, 505–531.

Howard, J. A. and J. N. Sheth (1969): *The Theory of Buyer Behavior*, Wiley.

Hsiao, C. (2014): *Analysis of panel data*, 54, Cambridge university press.

Khan, S., F. Ouyang, and E. Tamer (2019): "Inference on Semiparametric Multinomial Response Models," Working Paper.

Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

Li, L. (2019): "Identification of Structural and Counterfactual Parameters in a Large Class of Structural Econometric Models," Working Paper.

Luarn, P. and H.-H. Lin (2003): "A Customer Loyalty Model for E-Service Context," *Journal of Electronic Commerce Research*, 4, 156–167.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

Manski, C. F. and T. S. Thompson (1986): "Operational Characteristics of Maximum Score Estimation," *Journal of Econometrics*, 32, 85–108.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, 105–142.

Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, Elsevier, vol. IV, chap. 36.

Pakes, A. and J. Porter (2016): "Moment Inequalities for Multinomial Choice with Fixed Effects," Working Paper, National Bureau of Economic Research.

Powell, J. L. and P. A. Ruud (2008): "Simple Estimators for Semiparametric Multinomial Choice Models," *University of California, Berkeley*.

Reichheld, F. F. and P. Schefter (2000): "E-loyalty: Your Secret Weapon on The Web," *Harvard Business Review*, 78, 105–113.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Wasserman, L. (2006): *All of Nonparametric Statistics*, Springer Science & Business Media.

# Supplementary Materials for Online Publication

## B Pairwise Time Homogeneity of Errors

As mentioned in Section 2.2, Assumption 3 is stronger than necessary, and our identification strategy carries over under the weaker Assumption 3', which requires that $\boldsymbol{\epsilon}_{it} \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{i,ts}, \mathbf{A}_i\right)$. To see why Proposition 1 still holds, consider:

$$
\begin{aligned}
&\mathbb{E}\left[y_{ijt} - y_{ijs} | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right] \\
&= \int \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \neq j} u\left(\overline{\delta}_k, \, A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{it} | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&\quad - \int \mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \epsilon_{ijs}\right) \geq \max_{k \neq j} u\left(\underline{\delta}_k, \, A_{ik}, \epsilon_{iks}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{is} | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&= \int \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k \neq j} u\left(\overline{\delta}_k, \, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_i | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&\quad - \int \mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k \neq j} u\left(\underline{\delta}_k, \, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_i | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&= \int \left[\begin{array}{c} \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k \neq j} u\left(\overline{\delta}_k, \, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \\ -\mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k \neq j} u\left(\underline{\delta}_k, \, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \end{array}\right] \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_i | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right)
\end{aligned}
$$

where $\overline{\boldsymbol{\delta}} = \overline{\mathbf{X}}\beta_0$, $\underline{\boldsymbol{\delta}} = \underline{\mathbf{X}}\beta_0$, and $\tilde{\boldsymbol{\epsilon}}_i$ denotes generic realizations of $\boldsymbol{\epsilon}_{it}$ and $\boldsymbol{\epsilon}_{is}$ conditional on $\mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ and $\mathbf{A}_i$. Notice that the second equality follows from the assumption that $\boldsymbol{\epsilon}_{it} \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{i,ts}, \mathbf{A}_i\right)$.

Again, if $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j$, the bracketed term in the last line of the displayed equation above must be nonpositive for all realizations of $\mathbf{A}_i$ and $\tilde{\boldsymbol{\epsilon}}_i$, so that $\mathbb{E}\left[y_{ijt} - y_{ijs} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \leq 0$ for all realizations of $\mathbf{A}_i$, which further implies that

$$
\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} | \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right] \middle| \, \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \leq 0.
$$

Taking the logical contraposition again gives Proposition 1.

# C Sufficient Conditions for Point Identification

In this section, we prove sufficient conditions for the point identification of $\beta_0$. For simplicity of notation, we fix $T = 2$.We first need to impose an assumption of strict multivariate monotonicity on the function $\psi_j$ defined in (5).

**Assumption 8** (Strict Monotonicity of $\psi_j$)**.** *For any realized* $\mathbf{A}_i$*, the function* $\psi_j(\cdot, \mathbf{A}_i) : \mathbb{R}^J \to \mathbb{R}$ *is strictly increasing, i.e., if* $\overline{\delta}_j > \underline{\delta}_j$ *for all* $j$*, then* $\psi\left(\overline{\boldsymbol{\delta}}, \mathbf{A}_i\right) > \psi\left(\underline{\boldsymbol{\delta}}, \mathbf{A}_i\right)$.

We note that Assumption 8 is implied by a stronger version of Assumption 1 together with an additional condition on the support of $u$ given $(\mathbf{X}_i, \mathbf{A}_i)$.

**Assumption 1'** (Strict Monotonicity of $u$)**.** $u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$ *is strictly increasing in the index* $\delta_{ijt}$*, for every realization of* $(A_{ij}, \epsilon_{ijt})$.

**Assumption 1"** (Overlapping Supports)**.** *Conditional on any realization of* $\mathbf{X}_i$ *and* $\mathbf{A}_i$*, we have* $\bigcap_{j=1}^{J} int\left(Supp\left(u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)\right)\right) \neq \emptyset$.

In particular, Assumption 1" is directly implied by the assumption of $Supp\left(u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right)\right) = \mathbb{R}$ conditional on any realization of $\mathbf{X}_i$ and $\mathbf{A}_i$, which is again satisfied in additive panel multinomial choice models with scalar fixed effects a la $u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right) = X_{ijt}'\beta_0 + A_{ij} + \epsilon_{ijt}$ under the assumption of $Supp\left(\epsilon_{ijt} \middle| \mathbf{X}_i, \mathbf{A}_i\right) = \mathbb{R}$ as commonly imposed in the literature.

**Lemma 5.** *Assumptions 1' and 1" imply Assumption 8.*

Finally, we impose the following assumption on $\Delta\mathbf{X}_i$, with $\Delta X_{ij} := X_{ij1} - X_{ij2}$ for all individual $i$ and product $j$ across period 1 and period 2.

**Assumption 9** (Full-Directional Support of $\Delta\mathbf{X}_i$)**.** *Suppose either (a) or (b) is true:*

*(a)* $\mathbf{0} \in int\left(Supp\left(\Delta\mathbf{X}_i\right)\right)$.

*(b) There exists some* $k \in \{1, ..., d_x\}$ *such that* $\beta_0^k \neq 0$ *and* $Supp\left(\Delta X_{ij}^k \middle| \Delta X_{il},\, l \neq j\right) = \mathbb{R}$ *for all* $j \in \{1, ..., J\}$*. Furthermore, for all* $j \in \{1, ..., J\}$*,* $Supp\left(\Delta X_{ij} \middle| \Delta X_{il},\, l \neq j\right)$ *is not contained in a proper linear subspace of* $\mathbb{R}^{d_x}$.

Assumption 9(a) is satisfied when $(X_{ij})$ is continuous random vector. On the other hand, Assumption 9(b) can accommodate discrete regressors generally, but requires one continuous covariate with large support. Assumption 9 ensures that $\Delta X_{ij}^{'}\beta_0 > 0$ and $\Delta X_{ik}^{'}\beta_0 < 0$ for all $k \neq j$ hold simultaneously with strictly positive probability.

**Theorem 3** (Point Identification)**.** *Under Assumptions 2, 3, 8 and 9, $\beta_0$ is point identified on $\mathbb{S}^{D-1}$.*

*Proof.* Recall first that

$$\gamma_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \int \left[\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) - \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_i = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right).$$

Hence, under Assumption 8, we have

$$\overline{\delta}_j < \underline{\delta}_j \text{ and } \overline{\delta}_k > \underline{\delta}_k \text{ for all } k \neq j \quad \Rightarrow \quad \gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0, \tag{22}$$

since $\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) < \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)$ for every realization of $\mathbf{A}_i$. Together with Assumption 9, we deduce that

$$\mathbb{P}\left\{\gamma_{j,t,s}\left(\mathbf{X}_i\right) > 0\right\} \geq \mathbb{P}\left\{\Delta X_{ij}^{'}\beta_0 > 0 \ \wedge \ \Delta X_{ik}^{'}\beta_0 < 0, \forall k \neq j\right\} > 0.$$

Now for any $\beta \in \mathbb{S}^{D-1}\backslash\left\{\beta_0\right\}$, define for any product $j$,

$$H_j\left(\beta\right) := \left\{\mathbf{v} \in Supp\left(\Delta \mathbf{X}_i\right) : \ v_j^{'}\beta < 0 < v_j^{'}\beta_0, \ \wedge \ v_k^{'}\beta_0 < 0 < v_k^{'}\beta, \ \forall k \neq j\right\}.$$

As $\beta \neq \beta_0$, by Assumption 9 we know that

$$\mathbb{P}\left(\Delta \mathbf{X}_i \in H_j\left(\beta\right)\right) > 0. \tag{23}$$

Moreover, for any realization of $\mathbf{X}_i$ s.t. $\Delta \mathbf{X}_i \in H_j\left(\beta\right)$, we must have: (i) $\gamma_{j,t,s}\left(\mathbf{X}_i\right) > 0$ by (22), and (ii):

$$\lambda_j\left(\Delta \mathbf{X}_i, \beta\right) = \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}} \Delta X_{ik}^{'}\beta \geq 0\right\} = 1$$

so that $G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta \mathbf{X}_i, \beta\right) = G\left(\gamma_j\left(\mathbf{X}_i\right)\right) > 0$ for all such $\mathbf{X}_i$. Hence,

$$\mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right) \middle| \Delta \mathbf{X}_i \in H_j\left(\beta\right)\right] > 0. \tag{24}$$

Combining (23) and (24), we have:

$$\begin{aligned} Q_j(\beta) &= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta\mathbf{X}_i,\beta\right)\right] \\ &\geq \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta\mathbf{X}_i,\beta\right)\mathbb{1}\left\{\Delta\mathbf{X}_i \in H_j(\beta)\right\}\right] \\ &= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\mathbb{1}\left\{\Delta\mathbf{X}_i \in H_j(\beta)\right\}\right] \\ &= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\middle|\,\Delta\mathbf{X}_i \in H_j(\beta)\right]\mathbb{P}\left(\Delta\mathbf{X}_i \in H_j(\beta)\right) \\ &> 0 = Q_j(\beta_0). \end{aligned}$$

□

# D Additional Simulation Results

## D.1 Adaptive-Grid Computation Algorithm

In this section, we illustrate a typical output of our second-step computation algorithm based on the adaptive-grid search over the angle space, and show that the algorithm works well. For this purpose we consider a simplified DGP without fixed effect $A_{ij}$. We draw each of $X_{ijt}^{(d)}$ independently across each dimension $d \in \{1, ..., D\}$ from the standard normal distribution, and set the distribution of the idiosyncratic shock to be $\epsilon_{ijt} \sim TIEV(0,1)$, so that we can skip the first-step estimation and directly calculate the true conditional choice probability conditioned on each $\mathbf{X}_i$. Note that the conditions for point identification of $\beta_0$ are satisfied. Because we are only seeking to illustrate the validity of the algorithm itself, we set $N$ to be large with $N = 10^7$ and $D = 3, J = 3, T = 2$. Then we apply our adaptive-grid algorithm to search for $\beta_0$.

Figure 2 shows how our computational algorithm works in finding the true unknown $\theta_0$, the angle representation of the true $\beta_0$ in the $\Theta$ space. The horizontal and vertical axes correspond to the two polar coordinates that are associated with $\mathbb{S}^2$. The blue dots represent the points that our algorithm searches over but find *not* to be minimizers of the sample criterion $\hat{Q}$. The black box indicates the area that the minimizers for the sample criterion $\hat{Q}$ lie within, or more precisely, a rectangular enclosure of the numerical argmin set. The big black dot stands for the true parameter value $\theta_0 = (0.4205, 0.4636)'$.

It is evident from Figure 2 that our adaptive-grid algorithm is able to correctly

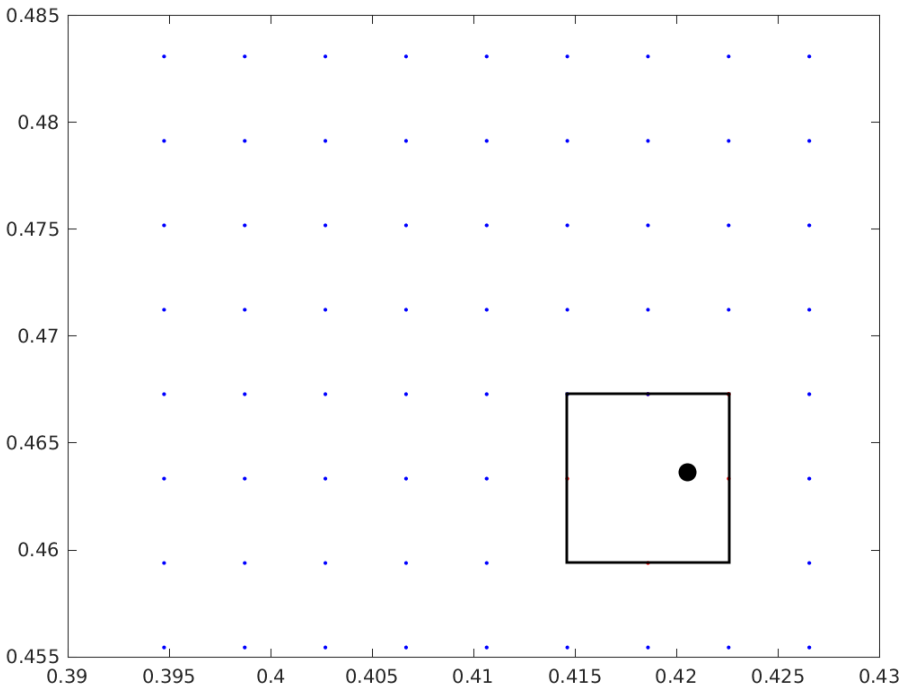



Figure 2: The Argmin Set in $\Theta$

locate an area that covers the true $\theta_0$, which lies within the small black box representing the estimated set of $\hat{\theta}$, demonstrating the efficacy of the algorithm. Besides, it is worth mentioning that our algorithm computes reasonably fast, as it first performs a rough search on the whole unit sphere $\mathbb{S}^2$, then focuses on the area where the minimizers are most likely to lie. In the last few rounds of search, the algorithm evaluates the criterion function $\hat{Q}$ on a relatively small area of points shown by those blue and red dots in Figure 2 until the desired level of accuracy is achieved.

For a more transparent representation, we translate the angles $\theta$ in the polar coordinates into unit vectors $\beta$ on the unit sphere $\mathbb{S}^2$ and show it in Figure 3.

Figure 3 is now plotted on $\mathbb{S}^2 \subseteq \mathbb{R}^3$. Again the blue dots represent the points that do not achieve the minimum of $\hat{Q}$; the black box shows an enclosing set of the minimizers of $\hat{Q}$. The big black dot represents the true parameter value $\beta_0$, which resides inside the black box of the minimizers of $\hat{Q}$. Figure 3 illustrates that our computation algorithm is able to locate a tight area around $\beta_0$.

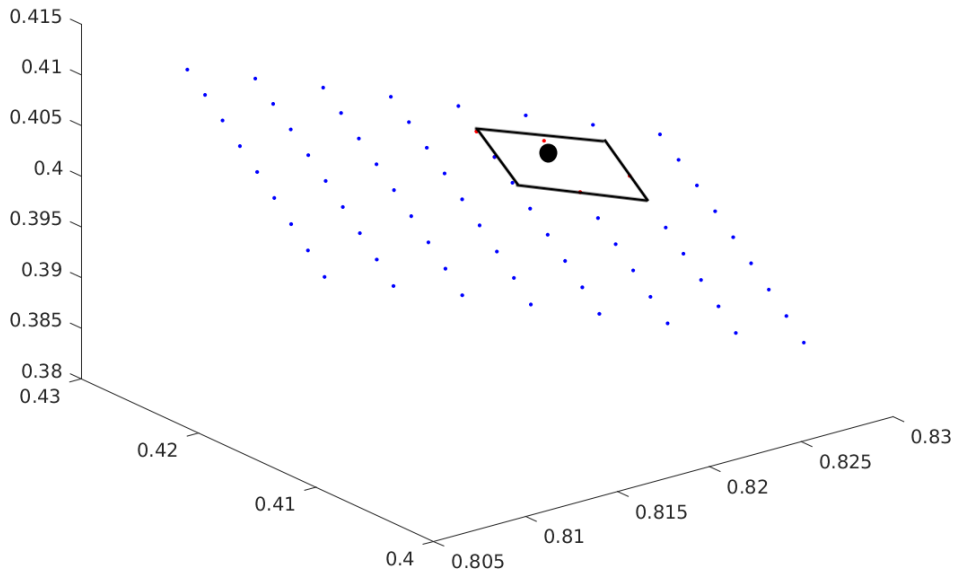



Figure 3: The Argmin Set in $\mathbb{S}^2$

## D.2 Results Varying $D, J, T$

In this section, we show how our estimator performs under different $(D, J, T)$. We maintain $N = 10,000$ as in the baseline configuration. We draw $Z_i \sim \mathcal{N}(0,1)$ and construct $A$ and $X$ according to the following specifications:

$$A_{ij} \sim \begin{cases} 0, & j = 1, \\ [Z_i]_+, & j = 2, \\ \mathcal{U}[-0.25, 0.25], & j = 3, ..., J, \end{cases} \qquad X_{ijt}^{(d)} \sim \begin{cases} U[-1,1], & d = 1, \\ Z_i + \mathcal{N}(0,6), & d = 2, \\ \mathcal{N}(0,1), & d = 3, ..., D, \end{cases}$$

which coincides with the baseline model at $D = 3, J = 3$. We emphasize that in all configurations we allow for nonlinear dependence between $A$ and $X$ via the latent variable $Z_i$.

We report in Table 9 the performance of our estimators for each of the corresponding configurations across all $M = 100$ simulations.

From Table 9 we find a larger $T$ improves the performance of our estimator, which is arguably more practically relevant given the increasing availability of long panel data nowadays. The improvement in performance with larger $T$ is because our method can extract more information from $T \times (T-1)$ ordered pairs of time

Table 9: Performance Varying $D, J, T$

| rMSE | $J=3$ | | $J=4$ | |
|---|---|---|---|---|
| | $T=2$ | $T=4$ | $T=2$ | $T=4$ |
| $D=3$ | 0.0745 | 0.0397 | 0.1137 | 0.0722 |
| $D=4$ | 0.0945 | 0.0580 | 0.1357 | 0.0807 |
| MND | $J=3$ | | $J=4$ | |
| | $T=2$ | $T=4$ | $T=2$ | $T=4$ |
| $D=3$ | 0.0648 | 0.0348 | 0.1005 | 0.0639 |
| $D=4$ | 0.0864 | 0.0539 | 0.1233 | 0.0750 |

periods which effectively increase the total number of observations. We also find that increase in $D$ or $J$ adversely affects the performance of our estimator, which is expected because more information is required to estimate more covariates ($D$) or deal with more alternatives ($J$). However, as can be seen from Table 9, the magnitude of such decline in performance is mild. For example, when $J$ is 4 and $T$ is 4, an increase in the dimension of product characteristics $D$ from 3 to 4 will increase the rMSE from 0.0722 to 0.0807. Likewise, when $D = 4$ and $T = 4$, an increase in $J$ from 3 to 4 will increase the rMSE from 0.0580 to 0.0807.

# Contents


**1 Introduction** **2**

**2 Panel Multinomial Choice Model** **7**

2.1 Model Setup . . . . . 7

2.2 Key Assumptions . . . . . 11

**3 Identification Strategy** **14**

**4 Estimation and Computation** **21**

4.1 A Consistent Two-Step Estimator . . . . . 21

4.2 Computation . . . . . 23

**5 Simulation** **27**

5.1 First-Stage Performance . . . . . 28

5.2 Two-Stage Performance . . . . . 29

**6 Empirical Illustration** **32**

6.1 Data and Methodology . . . . . 32

6.2 Results and Discussion . . . . . 34

6.3 A Possible Explanation via Monte-Carlo Simulations . . . . . 35

**7 Monotone Multi-Index Models** **37**

**8 Conclusion** **40**

**References** **42**

**B Pairwise Time Homogeneity of Errors** **46**

**C Sufficient Conditions for Point Identification** **47**

**D Additional Simulation Results** **49**

D.1 Adaptive-Grid Computation Algorithm . . . . . 49

D.2 Results Varying $D, J, T$ . . . . . 51