Robust Semiparametric Estimation in Panel Multinomial Choice Models*†

Wayne Yuan Gao‡ (*Job Market Paper*) and Ming Li§

January 31, 2019

Click here for the latest version.

Abstract

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 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 an index vector of observable characteristics, and uses the logical contraposition of an intertemporal inequality on choice probabilities to obtain identifying restrictions on the indexes. We provide consistent estimators based on our identification strategy, together with a computational procedure that exploits a combination of theoretical and practical advantages under a spherical-coordinate reparameterization. A simulation study and an empirical illustration with the Nielsen data are conducted to analyze the finite-sample performance of our estimation method and demonstrate the adequacy of our computational procedure for practical implementation.

Keywords: semiparametric estimation, panel multinomial choice, infinite-dimensional unobserved heterogeneity, nonseparability, monotonicity, spherical coordinates.

*We are grateful to Xiaohong Chen, Peter C.B. Phillips, and Philip A. Haile for their invaluable advice, guidance and encouragement. We thank Don Andrews, Isaiah Andrews, Benjamin Connault, Yuichi Kitamura, Patrick Kline, Matt Myung Hwan Seo, Xiaoxia Shi, Sheng Xu for helpful discussions and suggestions. 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.

‡Dept. of Economics, Yale University. 28 Hillhouse Ave., New Haven, CT 06511. wayne.gao@yale.edu.

§Dept. of Economics, Yale University. 28 Hillhouse Ave., New Haven, CT 06511. ming.li@yale.edu.

Contents

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, thus incorporating rich forms of unobserved heterogeneity. Our identification strategy exploits multivariate monotonicity in its contrapositive form, which provides powerful leverage for converting 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.

Our framework involves 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 β_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, while the last term ϵ_{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 ϵ_{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 ϵ_{ijt} is accommodated in addition, allowing for in particular unobserved 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, then by logical contraposition it *cannot* be possible that this product becomes *worse* while all other products become *better* across 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. Due to the intrinsic lack of scale identification induced by the discreteness in our model (and most discrete choice models), angles arise as a natural reparameterization of the parameter space under a normalization on the scale of the index parameter β_0 to be unity. When endowed

with the natural *great-circle metric*, the angle space becomes a parameter space that is both compact and convex, while preserving the spherical geometry among the observational equivalence classes in the original parameter space of β_0. We discuss why this combination of niceties enjoyed by our reparameterization brings about both theoretical and computational advantages relative to other forms of normalization or parameterization in the previous literature.

Our computation procedure then exploits the compactness and convexity of our angle parameter space, represented essentially in the form of a hyper-rectangle. We deploy a bisection-style *nested rectangle algorithm* that shrinks and refines an adaptive grid recursively to any chosen precision, with a technical adaption to account for the underlying spherical geometry. Moreover, our grid-based algorithm handles well the discreteness in our criterion functions, which renders usual gradient-based algorithms inapplicable.

A simulation study is then conducted to analyze the finite-sample performance of our method and the adequacy of our computational procedure for practical implementation. We investigate under different model configurations the performances of the first-stage machine learning estimators and the final estimators obtained through our second-stage computational algorithm, 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 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.

As an extension, the estimated model is shown to be further utilizable for counterfactual analysis, such as predicting the effect of a promotional campaign on product sales. We show that monotonicity in the parametric index structure provides a key lever to separate the direct effect of observable characteristics on choice probabilities from the indirect correlation effect between observable characteristics and unobserved heterogeneity. This separation allows us to predict the counterfactual effect of an exogenous change in observable characteristics, with the unobserved agent-product fixed effects held fixed, which can be achieved in a long panel setting through a nonparametric time series regression of individual choices on the index vectors of observable characteristics.

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. To elaborate, we provide a specific example 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 (2018).

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. Assuming additive separability between the parametric index of observable characteristics and a scalar-valued function of unobserved heterogeneity terms, they are able to focus on the scalar differences between the parametric indexes of different products in the panel multinomial choice setting, and use the first-order stochastic dominance implied by monotonicity to derive identifying moment inequalities. 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 (2017) 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). A recent paper by Chernozhukov, Fernández-Val, and Newey (2017) 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 on *monotone single-index models*. A related class of estimators that leverages univariate monotonicity, known as *maximum score* or *rank-order estimator*s, dates back to a series of original contributions by Manski (1975, 1985, 1987), and is 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 theoretical difference from the single-index setting, leading to a significantly different method of estimation relative to rank-order estimators. See Pakes and Porter (2016) and our appendix for more discussion on this difference.

Finally, our model and method are complementary to another class of models that fall into the framework of *invertible multi-index models*, using the terminology of Ahn, Ichimura, Powell, and Ruud (2018). The celebrated paper by Berry, Levinsohn, and Pakes (1995a) 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 specific context of demand estimation, the recent work by Ahn, Ichimura, Powell, and Ruud (2018) provides a high-level treatment of multi-index models based on invertibility. In comparison, the method proposed in our paper does not involve invertibility, but relies instead on monotonicity.

The rest of this paper is organized as follows. Section 2 introduces our main model specifications and assumptions, and 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 empirical illustration with the Nielsen data. Section 7 discusses extension and generalization of our method, 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 7.2 for a more general discussion about the wide applicability of our proposed methods.

Specifically, we consider the following discrete choice model, which essentially 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 among 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

  $$\delta_{ijt} := X_{ijt}^{'}\beta_0 \tag{2}$$

  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. We will thereafter refer to A_{ij} as the (ij-specific) *fixed effect*.
- ϵ_{ijt} is an ijt-specific unobserved error term of arbitrary dimensions, which captures time-idiosyncratic utility shocks to product j for agent i, or in other words, agent i's statistical uncertainty in decision making regarding the choice of product j.
- u is an unknown measurable real-valued function, interpreted as a *utility function* that aggregates the parametric index $X_{ijt}^{'}\beta_0$, the fixed effect A_{ij} and the error term ϵ_{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:

$$y_{ijt} = \mathbb{1}\left\{X_{ijt}^{'}\beta_0 + A_{ij} + \epsilon_{ijt} \geq \max_{k\in\{1,\ldots,J\}} X_{ikt}^{'}\beta_0 + A_{ik} + \epsilon_{ikt}\right\}, \tag{3}$$

Such models have been studied in recent work by Khan, Ouyang, and Tamer (2017) 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 (3) in the following form:

$$y_{ijt} = \mathbb{1}\left\{g_j\left(X_{ijt}, \beta_0\right) + f_j\left(A_{ij}, \epsilon_{ijt}\right) \geq \max_{k\in\{1,\ldots,J\}} g_k\left(X_{ikt}, \beta_0\right) + f_k\left(A_{ik}, \epsilon_{ikt}\right)\right\}, \tag{4}$$

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, notice that, though the dimensions of (A_{ij}, ϵ_{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 ϵ_{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.,

$$u_{ij}\left(X_{ijt}^{'}\beta_0, \epsilon_{ijt}\right) \equiv u\left(X_{ijt}^{'}\beta_0, A_{ij}, \epsilon_{ijt}\right). \tag{5}$$

Written in this form, our formulation allows for flexible time-invariant heterogeneity in how the index $X_{ijt}^{'}\beta_0$ affects agent i's utility from product j. In other words, given a fixed value

[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.

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 can be represented by (5).

It should be pointed out in particular that the fixed effect A_{ij} also effectively incorporates unobserved variations in the distributions of error terms ϵ_{ijt}. For example, if we assume that ϵ_{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 in particular 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 requirement for explicit specification of the functional forms.

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 (3) to model (1), rich forms of unmodeled randomness under (3) are automatically absorbed into the infinite-dimensional error term ϵ_{ijt}, which nests in itself all possible latent variables that affect utilities in some appropriate yet unspecified ways.[2] As a result, the assumption that ties occur with zero probabilities is effectively a much weaker restriction under our current

[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-dimensional error ϵ_{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.

model (1) than under model (3).

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 apply in our current model. For example, the recent method proposed by Shi, Shum, and Song (2018) utilizes cyclical monotonicity 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 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 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 $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{it})_{t=1}^{T}$.

The first assumption below imposes a monotonicity restriction on the utility function.

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

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 an integral part of our semiparametric model: monotonicity endows the index δ_{ijt} with an interpretation as an objective summary statistic for the direct effect of observable covariates on agent utilities. In other words, δ_{ijt} may be regarded 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 β_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 may be trivially satisfied by construction, such as the standard panel multinomial choice model (3). In Section 7.2, we provide more examples of parametric and semiparametric models featured by monotonicity in an index structure beyond the multinomial choice setting.

Next, we impose the standard assumption of cross-sectional random sampling.

Assumption 2 (Random Sampling)**.** $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \epsilon_i)$ *is i.i.d. across* $i \in \{1, ..., N\}$.

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 $(\Omega, \mathscr{F}, \mathbb{P})$.

It is now worth noting that the main part of this paper considers a *short panel* setting, 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. Section 7.1 provides further discussion on the additional capabilities of our method afforded in long panels, a setting with growing practical importance given the increasing availability of panel data.

Finally, we impose the following stationarity assumption on the distribution of error terms.

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.,*

$$\boldsymbol{\epsilon}_{it} | \left(\mathbf{X}_i, \mathbf{A}_i\right) \ \sim \ \mathbb{P}\left(\cdot | \, \mathbf{A}_i\right).$$

Assumption 3 as presented here is strictly stronger than necessary, 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.,

$$\boldsymbol{\epsilon}_{it} | \left(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i\right) \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i\right).$$

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 Assumption 3'), a statement conditioned on the arbitrarily dimensional fixed effect $\mathbf{A}_i$ in a fully flexible manner, automatically absorbs all possible *time-invariant* heterogeneity and dependence structures across and within $\mathbf{X}_{it} = (X_{ijt})_{j=1}^{J}$, $\mathbf{A}_i = (A_{ij})_{j=1}^{J}$ and $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^{J}$.

In particular, the allowance for arbitrary distributional dependence between observable covariates $\mathbf{X}_{it}$ and the fixed effect $\mathbf{A}_i$ incorporates automatically the effects of *time-invariant* variables, or more precisely the effects of *time-invariant components* of relevant variables, that affect choice behaviors in economically significant ways. 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 such 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 (1995b) 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). It is thus worth pointing out that our 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 (1995b), 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 (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, β_0, in Section 3 and construct estimators of β_0 in Section 4. For the identification of more sophisticated counterfactual parameters that involve other unknown components of model (1), see Section 7.1 for further analysis.

3 Identification Strategy

Assumption 3' (Pairwise Stationarity of Errors). The conditional distribution of $\boldsymbol{\epsilon}_{it}$ given $(\mathbf{X}_i, \mathbf{A}_i)$ is stationary over time t, i.e.,

$$\boldsymbol{\epsilon}_{it} | \, (\mathbf{X}_i, \mathbf{A}_i) \;\sim\; \mathbb{P}\left(\cdot | \, \mathbf{A}_i\right).$$

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 7.2 for more details.

Our key identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form. As a reminder, we start with an explicit statement of multivariate monotonicity in the definition below, 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 $\mathbb{R}^J$,

$$\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j \text{ for all } j = 1, ..., J \quad \Rightarrow \quad \psi\left(\overline{\boldsymbol{\delta}}\right) \leq \psi\left(\underline{\boldsymbol{\delta}}\right).$$

Remark 1 (Logical Contraposition of Multivariate Monotonicity). The following is logically equivalent to weak increasingness as defined in Definition 1: for any pair of $\left(\overline{\boldsymbol{\delta}}, \underline{\boldsymbol{\delta}}\right)$,

$$\psi\left(\overline{\boldsymbol{\delta}}\right) > \psi\left(\underline{\boldsymbol{\delta}}\right) \quad \Rightarrow \quad \text{NOT } \left\{\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j \text{ for all } j = 1, ..., J\right\}. \tag{6}$$

where " NOT" denotes the logical negation operator.

Remark 1 is tautologically true given Definition 1, which simply states the standard notion of weak monotonicity for a multivariate real function ψ. It is important to note that ψ is scalar-valued, so that the logical negation of the inequality $\psi\left(\overline{\boldsymbol{\delta}}\right) \leq \psi\left(\underline{\boldsymbol{\delta}}\right)$ simply becomes $\psi\left(\overline{\boldsymbol{\delta}}\right) > \psi\left(\underline{\boldsymbol{\delta}}\right)$, which will not be true if ψ is vector-valued, as the negation of an vector inequality is no longer a vector inequality in general. For this very reason, we leave an explicit negation sign in the right-hand side of (6), which contains a vector inequality written out elementwisely.

Our subsequent identification strategy will leverage heavily the simple contraposition of monotonicity (6), 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 (6). Finally, we use the contraposition of monotonicity to obtain the right-hand side of (6), which will translate into identifying restrictions on the parameter β_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 $(\mathbf{X}_{it}, \mathbf{A}_i)$:

$$\begin{aligned}
& \mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right] \\
= & \int \mathbb{1}\left\{u\left(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \neq j} u\left(X_{ikt}'\beta_0,\ A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\epsilon_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i\right) \\
= & \int \mathbb{1}\left\{u\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \neq j} u\left(\delta_{ikt},\ A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\epsilon_{ijt} \middle| \mathbf{A}_i\right) \\
=: & \ \psi_j\left(\delta_{ijt}, \left(-\delta_{ikt}\right)_{k \neq j}, \mathbf{A}_i\right)
\end{aligned} \tag{7}$$

where the second equality follows from the index definition $\delta_{ijt} = X_{ijt}'\beta_0$ and Assumption (3) (Stationarity of Errors), which enables us to write ψ_j without the time subscript t.

Clearly, the monotonicity of the utility function u in the index argument δ_{ijt} (Assumption (1)) translates into the multivariate monotonicity of the function ψ_j in the vector of indexes $\left(\delta_{ijt}, (-\delta_{ikt})_{k\neq j}\right)$:

Lemma 1. *For any given realization of* $\mathbf{A}_i$*, the function* $\psi_j(\cdot, \mathbf{A}_i) : \mathbb{R}^J \to \mathbb{R}$ *is weakly increasing.*

Notice that 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 δ_{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 ψ_j by averaging out $\mathbf{A}_i$.

Step 2: Construction of an observable inequality

Consider the following intertemporal difference in conditional choice probabilities:

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right] \tag{8}$$

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 $\underline{\boldsymbol{\delta}}$. 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 $k \neq j \;\Rightarrow\; \gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \leq 0$.

To see why Lemma 2 is true, rewrite $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ using the law of iterated expectations as follows:

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

$$= \mathbb{E}\left[\mathbb{E}\left[y_{ijt}|\,\mathbf{X}_{it}=\overline{\mathbf{X}},\mathbf{A}_i\right]-\mathbb{E}\left[y_{ijs}|\,\mathbf{X}_{is}=\underline{\mathbf{X}},\mathbf{A}_i\right]\Big|\,\mathbf{X}_{it}=\overline{\mathbf{X}},\mathbf{X}_{is}=\underline{\mathbf{X}}\right]$$
$$= \int\left[\psi_j\left(\overline{\delta}_j,\left(-\overline{\delta}_k\right)_{k\neq j},\mathbf{A}_i\right)-\psi_j\left(\underline{\delta}_j,\left(-\underline{\delta}_k\right)_{k\neq j},\mathbf{A}_i\right)\right]\mathrm{d}\mathbb{P}\left(\mathbf{A}_i|\,\mathbf{X}_{it}=\overline{\mathbf{X}},\mathbf{X}_{is}=\underline{\mathbf{X}}\right).$$

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:

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

which holds 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 $\mathbb{P}\left(\mathbf{A}_i|\,\mathbf{X}_{it}=\overline{\mathbf{X}},\mathbf{X}_{is}=\underline{\mathbf{X}}\right)$, potentially a hugely complicated probability measure considering the infinite dimensionality of $\mathbf{A}_i$ and the unrestricted forms of dependence between $\mathbf{A}_i$ and $\mathbf{X}_i$.

Step 3: Derivation of the key identifying restriction

We now take the logical contraposition of Lemma 2 and obtain the following proposition.

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

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)>0 \;\Rightarrow\; \text{NOT}\;\left\{\left(\overline{X}_j-\underline{X}_j\right)'\beta_0\leq 0 \text{ and } \left(\overline{X}_k-\underline{X}_k\right)'\beta_0\geq 0 \text{ for all } k\neq j\right\}. \tag{9}$$

Recall that $\delta_{ijt}=X_{ijt}'\beta_0$, so Proposition 1 follows immediately from Lemma 2 and defines an identifying restriction on β_0 that is free of all unknown nonparametric heterogeneity terms u, $\mathbf{A}$ and $\boldsymbol{\epsilon}$.

Proposition 1 is very intuitive given the monotonicity of conditional choice probabilities in the product indexes: if we observe an intertemporal increase in the conditional choice probability of product j from one period to another, then 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 despite 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 7.2 for a general framework under which the contraposition of monotonicity may be utilized. In particular, in a companion paper (Gao, Li, and Xu, 2018), we adapt this idea to the additional complications induced in a network formation setting, where nonseparability arises naturally from nontransferable utilities.

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 (9) in Proposition 1 by

$$\lambda_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}; \beta\right) := \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k\neq j\}}\left(\overline{X}_k - \underline{X}_k\right)' \beta \leq 0\right\}, \tag{10}$$

where $(-1)^{\mathbb{1}\{k\neq j\}}$ takes the value -1 for $k \neq j$ and the value 1 for $k = j$. Therefore, Proposition 1 can be written algebraically as

$$\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\} = 1 \quad \Rightarrow \quad \lambda_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}; \beta_0\right) \equiv 0,$$

for any possible realization $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$.

We may now define the following criterion function by taking a cross-sectional expectation over the random realization of $(\mathbf{X}_{it}, \mathbf{X}_{is})$:

$$\begin{aligned} Q_{j,t,s}(\beta) &:= \mathbb{E}\left[\mathbb{1}\left\{\gamma_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right) > 0\right\} \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right)\right], \\ &\geq 0 = Q_{j,t,s}(\beta_0) \end{aligned} \tag{11}$$

which is clearly minimized to zero at the true parameter value β_0. Without normalization and further assumptions for point identification, there might be multiple values of β_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.,

$$G(z) \begin{cases} > 0, & \text{if } z > 0, \\ = 0, & \text{if } z \leq 0. \end{cases} \tag{12}$$

we may define $Q^G_{j,t,s}$ by the following:

$$Q^G_{j,t,s}(\beta) := \mathbb{E}\left[G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)\right) \ \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right)\right], \tag{13}$$

which is also clearly minimized to zero at the true parameter value β_0.

The sign-preserving function G, if also set to be monotone, continuous or bounded, may 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 analogs of the population criterion function defined here. It is worth pointing out, however, that this smoothing function G is built into the *population* criterion function as in (13), 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 $Q_{j,t,s}^{G}$ and simply write $Q_{j,t,s}$ throughout this paper, except where the functional forms of G become significant for some particular results.

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:

$$Q(\beta) := \sum_{j=1}^{J} \sum_{t \neq s}^{T} Q_{j,t,s}(\beta), \quad \text{for any } \beta \in \mathbb{R}^{D}, \tag{14}$$

which is again minimized to zero at the true parameter value β_0.

Essentially, our criterion function is constructed to be an aggregation of the identifying restrictions on β_0 in the form of discrete Boolean variables across all (j, t, s) in the data, 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 (11) and (13), 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 not possible to push inside (or outside) 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 ($J = 1$)[3], in which our population criterion function essentially degenerates to the maximum-score or rank-order criterion function, as can be easily seen from the following derivations, where G is taken to be the positive part function $G(z) = [z]_+$, the product subscript $j = 1$ is suppressed, and X_t now simply denotes the *vector* of observable covariates:

$$\begin{aligned}
&Q_{t,s}(\beta) + Q_{s,t}(\beta) \\
=&\mathbb{E}\left[[\gamma(X_t, X_s)]_+ \mathbb{1}\{(X_t - X_s)\beta \geq 0\}\right] + \mathbb{E}\left[[\gamma(X_s, X_t)]_+ \mathbb{1}\{(X_s - X_t)\beta \geq 0\}\right] \\
=&\mathbb{E}\left[[\gamma(X_t, X_s)]_+ \operatorname{sgn}((X_t - X_s)\beta)\right] + \mathbb{E}\left[[-\gamma(X_t, X_s)]_+ [-\operatorname{sgn}((X_t - X_s)\beta)]\right] \\
=&\mathbb{E}\left[\left([\gamma(X_t, X_s)]_+ - [-\gamma(X_t, X_s)]_+\right) \operatorname{sgn}((X_t - X_s)\beta)\right] \\
=&\mathbb{E}\left[\gamma(X_t, X_s) \operatorname{sgn}((X_t - X_s)\beta)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[y_t - y_s | X_t, X_s\right] \operatorname{sgn}((X_t - X_s)\beta)\right] \\
=&\mathbb{E}\left[(y_t - y_s) \operatorname{sgn}((X_t - X_s)\beta)\right], \qquad (15)
\end{aligned}$$

The last line (15) is a familiar maximum-score or rank-order criterion function, constructed based on an equivalence relationship induced by univariate monotonicity of the following form:

$$\gamma(X_t, X_s) > 0 \;\Leftrightarrow\; (X_t - X_s)\beta > 0, \qquad (16)$$

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

$$\gamma(X_t, X_s) > 0 \Rightarrow \text{NOT}\{(X_t - X_s)\beta \leq 0\} \Leftrightarrow (X_t - X_s)\beta > 0 \Rightarrow \gamma(X_t, X_s) \geq 0, \qquad (17)$$

If weak monotonicity is strengthened to strict monotonicity, (17) immediately gives (16).

However, such equivalence relationships as (16) or (17) *cannot* be generalized to the multivariate setting with $J \geq 2$, as the right hand side of (9),

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

does not imply $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \geq 0$ in the converse direction. This breaks the bidirectionality built into the maximum-score or rank-order criterion function, and thus we can no longer

[3]This arises naturally in binomial choice models with the characteristics of the outside option set to be zero completely. 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.

aggregate $Q_{j,t,s}$ and $Q_{j,s,t}$ from two opposite directions into a unified representation as in (15). See a more general discussion on this difference in Appendix D.

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

We may now present our set identification result.

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

$$\beta_0 \in B_0 := \left\{\beta \in \mathbb{R}^D : Q(\beta) = 0\right\}. \tag{18}$$

We will refer to B_0 as the *identified set*.

In Appendix C, we provide sufficient conditions for point identification of β_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 Khan, Ouyang, and Tamer (2017) and Shi, Shum, and Song (2018).[4]

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 in the seminal paper by Manski (1975). Of course, whenever the additional assumptions for point identification are satisfied in data, the set estimator will automatically shrink to a point asymptotically.

4 Estimation and Computation

4.1 Two-Stage Estimation Procedure

We now proceed to construct consistent estimators of the identified set B_0 defined in (18), and provide a computation procedure designed to exploit the niceties and in the meanwhile

[4]It might be worth pointing out that the identification arguments in Khan, Ouyang, and Tamer (2017) and Shi, Shum, and Song (2018) 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), do not involve conditioning on *equalities*, but only *inequalities* that define (intersections of) half-spaces in the parameter space $\mathbb{R}^D$.

confront the challenges in our estimation problem.

We construct our estimator in the framework of extremum estimation, and start by defining the sample criterion functions. For each product j and each pair of periods (t, s), we define the sample analog of the population criterion function $Q_{j,t,s}$, as defined in (13), by the following:

$$\hat{Q}_{j,t,s}(\beta) := \frac{1}{N}\sum_{i=1}^{N} G\left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)\right) \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right) \tag{19}$$

where $\hat{\gamma}_{j,t,s}$ is some chosen nonparametric estimator of $\gamma_{j,t,s}$ and G is some chosen one-sided sign preserving function. This leads to the implied dependence of the final estimate $\hat{\beta}$ on both the choice of $\hat{\gamma}$ and the choice of G, which we will discuss in more details later.

Our estimator may naturally be obtained through a two-stage procedure.

The first stage takes the form of a standard nonparametric regression, in which we obtain a nonparametric estimator $\hat{\gamma}_{j,t,s}$ of $\gamma_{j,t,s}$ for each (j, t, s). Applied researchers could choose from an abundance of kernel-based or sieve-based estimators. In this paper, we adopt a machine learning algorithm using single-layer neural network sieves. Section (4.2) provide more in-depth discussion about the first stage.

The second stage of our estimation procedure solves an optimization problem, i.e., to numerically compute minimizers of the aggregated sample criterion function:

$$\hat{Q}(\beta) := \sum_{j=1}^{J}\sum_{t\neq s}^{T} \hat{Q}_{j,t,s}(\beta).$$

In our implementation, we first impose in Section (4.3.1)a scale normalization on the parameter of interest β_0 by restricting its Euclidean norm to unity, effectively transforming the parameter space into the unit sphere in $\mathbb{R}^D$. Moreover, we further reparameterize the unit sphere in polar coordinates, or more precisely *angles*, which enjoy a combination of topological, geometric and arithmetic advantages, such as compactness and convexity. Section (4.3.2) provides consistency results for our set estimators in the Hausdorff set distance, while consistency under further assumptions for point identification is established in Appendix C. Lastly, Section (4.3.3) contains more detailed discussions about our computation procedure, which exploits the compactness and convexity of our reparameterized parameter space and utilizes a bisection-style algorithm that recursively shrink and refine to a conservative enclosure of the minimizers with arbitrarily chosen precision.

4.2 First Stage: Nonparametric Regression

The first stage of our procedure concerns with estimating the intertemporal differences in conditional choice probabilities of the following form, as derived in (8):

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

for all on-support realizations $\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$, all pair of periods (t,s) and all products j. The estimation of $\gamma_{j,t,s}$ boils down to a nonparametric regression of the following form

$$\text{regress }\left(y_{ijt}-y_{ijs}\right)\text{ on }\left(\text{vec}\left(\mathbf{X}_{it}\right)',\text{vec}\left(\mathbf{X}_{is}\right)'\right)\text{ across all agents }i=1,...,N. \tag{20}$$

Many kernel-based and sieve-based methods have been developed in statistics and econometrics, with potentially different properties developed under various conditions. For example, see Chen (2007) and Wasserman (2013) for more comprehensive surveys of methods on nonparametric estimation. Hence in this paper we will simply take as given the results available in the literature on the first-stage nonparametric regression.

Specifically, 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 regression (20) includes time-varying observable characteristics of all products from two periods, the potentially high dimensionality of the regression 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, which we will use to derive the consistency of our final estimators obtained in the second stage. 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), who provides a front end for cross validation and hyperparameter tuning.

As our identification strategy is based on the logical implication of the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)>0$, we are intrinsically only interested in estimating whether the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)>0$ occurs, that is, a *binary* functional of $\gamma_{j,t,s}$ in the form of

$$\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)>0\right\},$$

but we are not interested in the exact magnitude of $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$.

When $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$ is close to zero, the estimation of $\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)>0\right\}$ may not be very precise relatively. Yet 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 $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$. We can exploit these intuitions by setting the one-sided sign-preserving function G, defined in (12), to be both Lipschitz continuous and bounded above, so that observations for $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$ close to zero are down-weighed in $Q_{j,t,s}$ via $G\left(\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)\right)$, while variations for positive and large $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$ are dampened in the same time.

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 due to the fact that 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 due to the fact that $\gamma_{j,t,s}=-\gamma_{j,s,t}$ by construction, so we may estimate either (t,s) or (s,t).[5]

It might be worth noting that our first-stage nonparametric estimation does not involve the parameter β_0, so there is no need to run the machine learning algorithm for different candidate parameter value β. Moreover, as all economic structures are supplied in the second stage through the criterion function we defined earlier, the use of machine learning algorithms does not interfere at all with the economic interpretability of our final estimates.

4.3 Second Stage: Extremum Estimation

4.3.1 Normalization and Reparameterization

The second stage of our estimation procedure concerns with the numerical minimization of the sample criterion function:

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

Before we solve this minimization problem, however, we first discuss about the inherent indeterminacy of the scale of β_0 in our model, and impose a particular form of scale normalization, followed by a reparamterization, to deal with this indeterminacy without affecting the implicit topological and metric structures in our problem.[6]

[5] Notice, however, that each ordered pair (t,s) or (s,t) provides complementary identifying information, as $\lambda(\mathbf{X}_{it},\mathbf{X}_{is};\beta)$ and $\lambda(\mathbf{X}_{is},\mathbf{X}_{it};\beta)$ do not admit such kind of deterministic relationship.

[6] The reason why we did not discuss about this scale indeterminacy in earlier sections is because our previous set identification results automatically recognize and accommodate this indeterminancy. Given that normalization is merely a representational device to index the underlying equivalence classes in the parameter space, its significance is conceptually more relevant in the implemention of estimation procedures

We begin with a discussion on observational equivalence relations for our main model (1). Clearly, for any positive constant $c > 0$, we may redefine the unknown parameter β_0 and the unknown utility function u in the following way:

$$\overline{\beta}_0 := \frac{1}{c}\beta_0, \quad \overline{u}\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right) := \overline{u}\left(c \cdot \delta_{ijt}, A_{ij}, \epsilon_{ijt}\right), \text{ for all } ijt.$$

Clearly, model (1) as characterized by $(\beta_0, u, \mathbf{A}, \boldsymbol{\epsilon}, \mathbf{X}, \mathbf{y})$ is observationally equivalent to the reparameterized model characterized by $\left(\overline{\beta}_0, \overline{u}, \mathbf{A}, \boldsymbol{\epsilon}, \mathbf{X}, \mathbf{y}\right)$, because by construction:

$$\overline{u}\left(X_{ijt}^{'}\overline{\beta}_0, A_{ij}, \epsilon_{ijt}\right) \equiv u\left(X_{ijt}^{'}\beta_0, A_{ij}, \epsilon_{ijt}\right).$$

If $\beta_0 \neq \mathbf{0}$, then we could take $c = \|\beta_0\|$ and normalize

$$\overline{\beta}_0 := \frac{1}{\|\beta_0\|}\beta_0 \in \mathbb{S}^{D-1},$$

where $\mathbb{S}^{D-1} := \left\{v \in \mathbb{R}^D : \|v\| = 1\right\}$ denotes the unit sphere in $\mathbb{R}^D$.

Applied to the identified set B_0 defined in (18), this scale normalization essentially imposes the following restriction:

$$\overline{B}_0 \subseteq \mathbb{S}^{D-1}. \tag{21}$$

We now make a stronger claim about the validity of (21), even when $\beta_0 = \mathbf{0}$.

Proposition 2 (Scale Normalization WLOG)**.** *The scale normalization* (21) *is without loss of generality (WLOG) in terms of the represented observational equivalence classes defined on the unknown components of model* (1)*.*

To see this, we focus on the potentially pathological case of $\beta_0 = \mathbf{0}$, which implies that $\delta_{ijt} \equiv 0$ for all ijt. Then, $(\beta_0, u, \mathbf{A}, \boldsymbol{\epsilon}, \mathbf{X}, \mathbf{y})$ is observationally equivalent to $\left(\overline{\beta}_0, \overline{u}, \mathbf{A}, \boldsymbol{\epsilon}, \mathbf{X}, \mathbf{y}\right)$ for any $\overline{\beta}_0 \in \mathbb{S}^{D-1}$ and

$$\overline{u}\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right) := \overline{u}\left(0, A_{ij}, \epsilon_{ijt}\right),$$

because again we have

$$\overline{u}\left(X_{ijt}^{'}\overline{\beta}_0, A_{ij}, \epsilon_{ijt}\right) \equiv u\left(X_{ijt}^{'}\beta_0, A_{ij}, \epsilon_{ijt}\right).$$

This reflects the simple fact that any degeneracy with respect to observable covariates can be equivalently induced by the degeneracy of index parameter β_0 or the degeneracy of the

rather than in identification analysis.

utility function u with respect to the index argument, and observable data cannot distinguish one from the other. Hence, we might as well use the nonparametric utility function u to absorb this denegeracy and normalize the scale of β_0 to unity.

Translated to our identification results, $\beta_0 = \mathbf{0}$ or degeneracy of u in its first argument imply that $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \equiv 0$ and thus $Q(\beta) \equiv 0$, giving a trivial identified set $B_0 = \mathbb{S}^{D-1}$, which is consistent with (21) and in the same time admits a clear interpretation: the identification result is agnostic about the direction of β_0, as there is no (populational) variation in the data to begin with.

The normalized parameter space $\mathbb{S}^{D-1}$ is a compact and connected space, but it is not convex under standard vector arithmetic. We now adopt a natural reparameterization of the unit sphere that preserves all its topological and geometric niceties but simultaneously brings about convexity, which will turn out to be handy for the implementation of our computation algorithm.

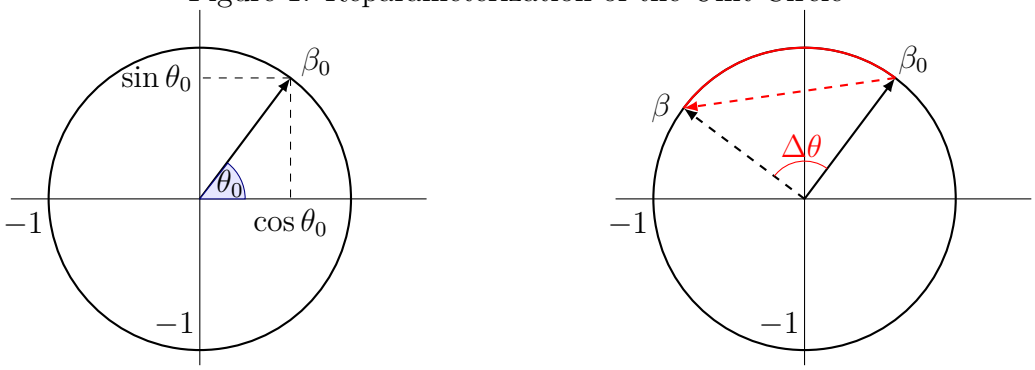
To fix ideas, we start with two examples, the unit circle $\mathbb{S}^1$ and the unit 2-sphere $\mathbb{S}^2$, in which case our reparameterization closely relates to familiar concepts such as polar coordinates.

Example 1 (Unit Circle $\mathbb{S}^1$ in $\mathbb{R}^2$)**.** Let $D = 2$, in which case $\mathbb{S}^{D-1}$ reduces to the unit circle as illustrated in Figure 1. In this case, $\Theta = [-\pi, \pi)$ and a point $\beta_0 \in \mathbb{S}^1$ can be represented by $\theta_0 \in \Theta$ via a mapping ω defined by:

$$\beta_0 = \omega(\theta_0) := \begin{pmatrix} \cos\theta_0 \\ \sin\theta_0 \end{pmatrix}.$$

There are at least two straightforward ways to measure the distance between two points on the unit circle, say, β_0 and β. First, we can use the standard Euclidean metric on $\mathbb{R}^2$, represented by the length of the straight line segment connecting β_0 and β in Figure1. Second, we can use the *great-circle metric* on $\mathbb{S}^1$, represented by the length of the inferior arc connecting β_0 and β, which is exactly equal to the radian of the angle $\Delta\theta$ between the two vectors β_0 and β. It turns out that the Euclidean metric and the great-circle metric are

Figure 1: Reparameterization of the Unit Circle



strongly equivalent[7] in consideration of the following relationship:

$$\|\beta - \beta_0\| = 2\sin\left(\frac{1}{2}\Delta\theta\right) \leq \Delta\theta \leq \frac{\pi}{2}\|\beta - \beta_0\|,$$

with $\|\beta - \beta_0\|$ bounded in $[0, 2]$ and $\Delta\theta$ bounded in $[0, \pi]$. Consequently, the choice of either metric leads to no differences in all topological structures and most essential metric structures such as uniform continuity and convergence. □

Example 2 (Unit Sphere $\mathbb{S}^2$ in $\mathbb{R}^3$)**.** Let $D = 3$. $\mathbb{S}^2$, the standard unit sphere, is illustrated in Figure 2. In this case, $\Theta = [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and a transformation mapping $\omega : \Theta \to \mathbb{S}^2$ can be defined by

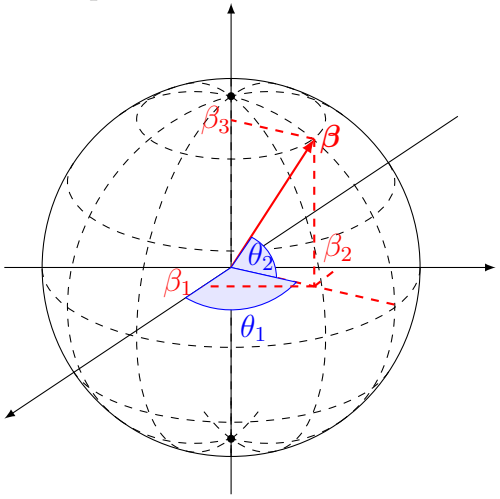
$$\beta = \omega(\theta) := \begin{pmatrix} \cos\theta_2 \cos\theta_1 \\ \cos\theta_2 \sin\theta_1 \\ \sin\theta_2 \end{pmatrix},$$

It should be noted that, for the case of $D \geq 3$, there exist many valid ways to define Θ and the transformation mapping ω. The one defined above is just one illustrative example, which exactly corresponds to the widely used *longitude-latitude* coordinate system on the surface of Earth. In particular, notice that an asymmetry is generated between the first and the remaining coordinates of $\theta \in \Theta$: while θ_1 can vary from $-\pi$ to π with half-closed half-open boundaries, the domain of every other coordinates is given by the closed interval $\left[-\frac{1}{2}\pi, \frac{1}{2}\pi\right]$, so that there exists a one-to-one mapping, i.e., ω, between Θ and $\mathbb{S}^{D-1}$. The intuition behind this asymmetry can be easily seen from the longitude-latitude coordinates: while longitude varies from 180 degree west to 180 degree east, latitude varies from 90 degree south to 90

[7] Two metrics ρ_1 and ρ_2 defined on the same point space $\mathcal{X}$ is said to be *strongly equivalent* if there exist two positive constants $\bar{c}, \underline{c}$ such that

$$\underline{c}\rho_1(x, y) \leq \rho_2(x, y) \leq \bar{c}\rho_1(x, y), \quad \forall x, y \in \mathcal{X}.$$

Figure 2: Reparameterization of the Unit Sphere



degree north. □

In general, for $D \geq 2$, we reparameterize $\mathbb{S}^{D-1}$ with $(D-1)$ angles in spherical coordinates. Specifically, define the angle space Θ by

$$\Theta := [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}, \tag{22}$$

and the transformation $\omega : \Theta \to \mathbb{S}^{D-1} \subseteq \mathbb{R}^D$ by the following mapping $\omega : \theta \mapsto \beta$:

$$\begin{cases} \beta_1 & := \cos\theta_{D-1} \ldots \cos\theta_2 \cos\theta_1, \\ \beta_2 & := \cos\theta_{D-1} \ldots \cos\theta_2 \sin\theta_1, \\ \vdots & \quad \vdots \\ \beta_{D-1} & := \cos\theta_{D-1} \sin\theta_{D-2}, \\ \beta_D & := \sin\theta_{D-1}, \end{cases} \tag{23}$$

which is a direct generalization of the transformation mappings in Examples 1 and 2.

We now endow Θ with a natural metric ρ_Θ that reflects its inherent spherical geometry, known as the *great-circle metric*, which we import from the natural *great-circle distance* defined on the unit sphere. Specifically, for any two points $\overline{\beta}, \underline{\beta}$ on the unit sphere $\mathbb{S}^{D-1}$, the great-circle distance $\rho_{GC}\left(\overline{\beta}, \underline{\beta}\right)$ between $\overline{\beta}$ and $\underline{\beta}$ is defined as the length of the inferior arc of the great circle that both $\overline{\beta}$ and $\underline{\beta}$ lie on. As the radius of the unit sphere is one by definition, $\rho_{GC}\left(\overline{\beta}, \underline{\beta}\right)$ is equal to the angle between the two vectors $\overline{\beta}$ and $\underline{\beta}$:

$$\rho_{GC}\left(\overline{\beta}, \underline{\beta}\right) := \arccos\left(\overline{\beta}' \underline{\beta}\right) \in [0, \pi], \quad \forall \overline{\beta}, \underline{\beta} \in \mathbb{S}^{D-1}.$$

Relative to the Euclidean distance in $\mathbb{R}^D$, the great-circle distance better reflects the curvature of $\mathbb{S}^{D-1}$, and has been widely used in other theoretical and applied disciplines such as physics, astronomy, geography and navigation. Considering now the angle space Θ, we import the great-circle distance ρ_{GC} to Θ via the transformation mapping ω, i.e.,

$$\rho_\Theta\left(\overline{\theta},\underline{\theta}\right) := \rho_{GC}\left(\omega\left(\overline{\theta}\right),\omega\left(\underline{\theta}\right)\right) = \arccos\left(\omega\left(\overline{\theta}\right)'\omega\left(\underline{\theta}\right)\right), \quad \forall\overline{\theta},\underline{\theta}\in\Theta. \tag{24}$$

Then the metric space (Θ,ρ_Θ) is by construction isometric to the metric space $\left(\mathbb{S}^{D-1},\rho_{GC}\right)$, so all topological and metric properties of $\left(\mathbb{S}^{D-1},\rho_{GC}\right)$ are preserved after the reparameterization.

Alternatively, we may import the standard Euclidean metric in $\mathbb{R}^D$ to the angle space Θ by defining, for all $\overline{\theta},\underline{\theta}\in\Theta$,

$$\rho_{Euc}\left(\overline{\theta},\underline{\theta}\right) := \left\|\omega\left(\overline{\theta}\right)-\omega\left(\underline{\theta}\right)\right\|. \tag{25}$$

Notice that, just as shown for the $D=2$ case in Example 1, the great-cricle metric ρ_Θ and the imported Euclidean metric ρ_{Euc} are *strongly equivalent*, considering that

$$\rho_{Euc}\left(\overline{\theta},\underline{\theta}\right) \leq \rho_{Euc}\left(\overline{\theta},\underline{\theta}\right) \leq \frac{\pi}{2}\rho_{Euc}\left(\overline{\theta},\underline{\theta}\right)$$

holds for any $D\geq 2$ in general. Hence, the choice between ρ_Θ and ρ_{Euc} is largely inconsequential in terms of topological and metric structures. However, the great-circle metric ρ_Θ naturally arises in the derivation of our results on consistency and convergence rates, so we will focus on ρ_Θ from now on.

The metric space (Θ,ρ_Θ) enjoys many niceties. First, while the unit sphere $\mathbb{S}^{D-1}$ is not convex, the new parameter space Θ is convex in the form of a $(D-1)$-dimensional hyper-rectangle, making it easy to take averages (or find bisection points) computationally in the parameter space. Second, (Θ,ρ_Θ) 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. Third, it also preserves the geometric structures of the sphere, including for instance the obvious observation that $-\pi$ and π in the first coordinate of Θ should be treated as exactly the same point, or more rigorously,

$$\rho_\Theta\left((\pi-\epsilon,\theta_2,...,\theta_{D-1}),(-\pi,\theta_2,...,\theta_{D-1})\right)\to 0, \quad \text{as } \epsilon\searrow 0.$$

This seemingly trivial property is nevertheless important in defining and interpreting whether certain parameter estimates converge asymptotically or not, and provide conceptual

foundations for subsequent asymptotic theories.

The nontriviality of the theoretical niceties afforded by our reparameterization can be partially seen from a comparison with a popular alternative form of normalization or reparameterization, where the scale of the first coordinate of β_0 is normalized to unity based on the assumption that $\beta_{0,1} \neq 0$. More precisely, if $\beta_{0,1} \neq 0$, then we may define

$$\tilde{\beta}_0 := \begin{pmatrix} \text{sgn}(\beta_{0,1}) \\ \beta_{0,2}/|\beta_{0,1}| \\ \vdots \\ \beta_{0,D}/|\beta_{0,1}| \end{pmatrix} \in \tilde{B} := \{1,-1\} \times \mathbb{R}^{D-1} \subseteq \mathbb{R}^D. \tag{26}$$

A convenient choice of metric on $\tilde{B}$ that has been often used in previous work is to simply adopt the restriction of the standard Euclidean norm ρ_{Euc} on $\mathbb{R}^D$ onto $\tilde{B}$. There are, however, at least three advantages of the metric space (Θ, ρ_Θ), or equivalently $\left(\mathbb{S}^{D-1}, \rho_{GC}\right)$, relative to $\left(\tilde{B}, \rho_{Euc}\right)$.

First, Θ or $\mathbb{S}^{D-1}$ represents a strictly larger point space: every point in $\tilde{B}$ can be represented by a point in Θ or $\mathbb{S}^{D-1}$, but not vice versa. Specifically, points in $\left\{\beta \in \mathbb{S}^{D-1} : \beta_{0,1} = 0\right\}$ are not represented in $\tilde{B}$, and thus ruled out a priori from subsequent analysis. Hence, the parameterization (Θ, ρ_Θ) or $\left(\mathbb{S}^{D-1}, \rho_{GC}\right)$ is preferable when there is no strong a priori knowledge on which coordinate of β must be nonzero.

Second, topologically $\left(\tilde{B}, \rho_{Euc}\right)$ is not compact, but (Θ, ρ_Θ) is compact. In particular, if compactness in the style of $\tilde{\beta}_0 \in \{1,-1\} \times [-M, M]^{D-1}$ is nevertheless assumed ad hoc in $\left(\tilde{B}, \rho_{Euc}\right)$, as often found in proofs of consistency for extremum estimators, then it is not only points with $\beta_{0,1} = 0$ are ruled out, but a neighborhood of points with $|\beta_{0,1}| < \frac{1}{M}$ are also ruled out a priori. However, as (Θ, ρ_Θ) is compact *globally* (as a whole parameter space) by construction, there is no need for further restrictive assumption on compactness.

Lastly, and perhaps most importantly, the metric ρ_{Euc}, when restricted to $\tilde{B}$, produces distortions in the definition and interpretation of convergence. To see this, consider the simple case of $D = 2$ with $\tilde{\beta}^{(M)} = (1, M)' \in \tilde{B}$. The limit Euclidean distance between $\tilde{\beta}^{(M)}$ and $-\tilde{\beta}^{(M)}$ as we take $M \to \infty$ is given by:

$$\rho_{Euc}\left(\tilde{\beta}^{(M)}, -\tilde{\beta}^{(M)}\right) = 2\sqrt{1 + (D-1)\,M^2} \to \infty$$

suggesting an apparent interpretation that the two points $\tilde{\beta}^{(M)}$ and $-\tilde{\beta}^{(M)}$ are moving farther and farther away from each other. Now, observe that $\tilde{\beta}^{(M)}$ and $-\tilde{\beta}^{(M)}$ correspond exactly

to the points $\overline{\theta}^{(M)}$ and $\underline{\theta}^{(M)}$ in Θ given by

$$\begin{aligned}\overline{\theta}^{(M)} &:= \omega^{-1}\left(\tilde{\beta}^{(M)} / \left\|\tilde{\beta}^{(M)}\right\|\right) \\ \underline{\theta}^{(M)} &:= \omega^{-1}\left(-\tilde{\beta}^{(M)} / \left\|\tilde{\beta}^{(M)}\right\|\right).\end{aligned}$$

To see this more rigorously, notice that the equivalence class of $\mathbb{R}^D$ represented by $\overline{\theta}^{(M)}$ or $\tilde{\beta}^{(M)}$ is exactly the same for every finite M:

$$\left\{\beta \in \mathbb{R}^D : \beta = r\omega\left(\overline{\theta}^{(M)}\right) \text{ for some } r > 0\right\} \equiv \left\{\beta \in \mathbb{R}^D : \beta = r\tilde{\beta}^{(M)} \text{ for some } r > 0\right\}.$$

However, as $M \to \infty$, we have

$$\rho_\Theta\left(\overline{\theta}^{(M)}, \underline{\theta}^{(M)}\right) \leq \rho\left(\overline{\theta}^{(M)}, \frac{\pi}{2}\right) + \rho\left(\underline{\theta}^{(M)}, \frac{\pi}{2}\right) \to 0,$$

suggesting the opposite interpretation that $\overline{\theta}^{(M)}$ and $\underline{\theta}^{(M)}$ should be viewed as converging to each other! Such a drastic distinction in the definition of convergence in the two different metric spaces will lead to qualitatively different asymptotic results in terms of consistency, convergence rates or asymptotic distributions, which are fundamentally dependent on the choice of metrics. We take the view that our choice of metric space (Θ, ρ_Θ), which is isometric with $\left(\mathbb{S}^{D-1}, \rho_{GC}\right)$, is more natural in its representation of the underlying equivalence classes induced by the lack of scale identification than the metric space $\left(\tilde{B}, \rho_{Euc}\right)$.

We have now presented our scale normalization, as well as the subsequent angle-space reparameterization that preserves important topological and geometric structures inherent in our model, and that brings about additional niceties that will be made clearer in both our convergence-rate results and our computational algorithm to be introduced in the next two subsections.

4.3.2 Consistency

As discussed in Section 4.2, we take as given the available results in the literature on the first-stage nonparametric regression, and state the the following assumption regarding first-stage convergence.

Assumption 4 (First-Stage Convergence)**.** *There exists a sequence of positive constants*

(c_N) *such that, for any* (j,t,s),

$$\begin{aligned}\left\|\hat{\gamma}_{j,t,s}-\gamma_{j,t,s}\right\|_2 &:= \sqrt{\int\left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)^2 \mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)} \\ &= O_p\left(c_N\right),\end{aligned}$$

and

$$N^{\frac{1}{2}}c_N \to \infty, c_N \to 0, \quad \text{as } N \to \infty.$$

For single-layer neural networks in particular, Chen and White (1999) establishes that

$$c_N = \left(\frac{\log N}{N}\right)^{\frac{1+2/(d+1)}{4(1+1/(d+1))}} = o_p\left(N^{-\frac{1}{4}}\right).$$

Next, we impose some smoothness condition on the one-sided sign preserving function G, which the first stage estimate $\hat{\gamma}$ is plugged into. As G can be chosen arbitrarily, the assumption below is not restrictive at all, but simply provides a guideline for the specification of G.

Assumption 5 (Nice Smoothing Function)**.** *The one-sided sign preserving function* $G : \mathbb{R} \to \mathbb{R}_+$ *is also Lipschitz continuous and strictly increasing.*

The next assumption, Assumption 6, imposes regularity conditions on the data distribution, so that the population criterion function $Q(\theta)$ is continuous.

To state Assumption 6, we first define the following notations: for each realized pair of observable covariates $(\mathbf{X}_{it},\mathbf{X}_{is}) = \left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$ from two periods, define the following spherical-coordinate reparametrization:

$$\begin{aligned} r_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right) &:= \left\|\overline{X}_k - \underline{X}_k\right\| \\ v_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right) &:= \begin{cases}\left(\overline{X}_k-\underline{X}_k\right)/r_{ik}, & r_{ik}>0, \\ 0, & r_{ik}=0.\end{cases}\end{aligned}$$

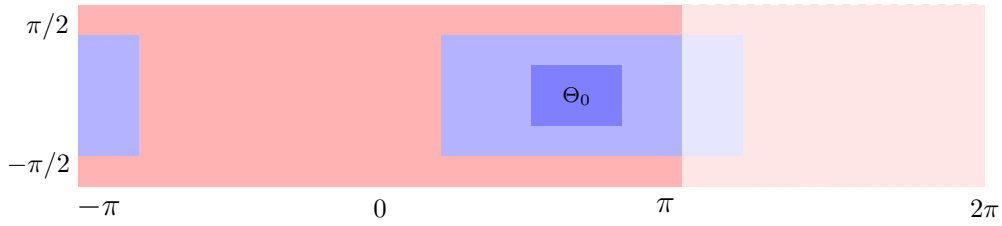
so that

$$\overline{X}_k - \underline{X}_k \equiv r_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right)\cdot v_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right),$$

i.e., $v_k\left(\overline{\mathbf{X}}-\underline{\mathbf{X}}\right)$ represents the direction of intertemporal change in observable characteristics of each product.

Assumption 6 (Regularity Condition)**.** *The distribution of* $v_k\left(\mathbf{X}_{it}-\mathbf{X}_{is}\right)$ *has no mass point for each* (k,t,s).

Figure 3: An Adaptive-Grid Algorithm



Assumption 6 is a fairly weak assumption: it essentially requires that the *directions* of intertemporal differences in observable characteristics are continuously distributed on their own supports. This allows all but one dimension of observable characteristics to be discrete.

Theorem 2 (Consistency)**.** *Under Assumptions 1-6, we have*

$$\rho_{\Theta,Hausdorff}\left(\hat{\Theta},\Theta_0\right) \xrightarrow{p} 0.$$

The asymptotic properties of the implied optimizer $\hat{\Theta}$, or correspondingly $\hat{B} = \omega\left(\hat{\Theta}\right)$, depend on the estimator $\hat{\gamma}$, the choice of G and the nature of the functional dependence induced by our nonstandard criterion function Q. Therein lies the difficulties in developing a full asymptotic theory of inference because of these induced complexities.

4.3.3 Computation Algorithm

Computationally, we search for minimizers of $\hat{Q}(\theta)$ in Θ exploiting the compactness and convexity of our spherical-coordinate parameter space Θ, which takes the form of a hyper-rectangle $\Theta = [-\pi,\pi) \times \left[-\frac{\pi}{2},\frac{\pi}{2}\right]^{D-2}$ as in (22).

Specifically, we 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. 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 maintain 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 3.

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

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 αth-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

$$\underline{\theta}_d^{(1)} := \min{}^* \underline{\Theta}_d^{(1)},$$
$$\overline{\theta}_d^{(1)} := \max{}^* \underline{\Theta}_d^{(1)},$$

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 recognize the periodicity of angles, i.e. $\theta_1 + 2\pi \equiv \theta_1$ and in particular $-\pi \equiv \pi$, as we move around the unit sphere (a Riemann surface). This is largely a programming nuisance: whenever $\underline{\Theta}_1^{(1)} \subsetneq \Theta_1^{(1)}$ crosses over at $-\pi$ and π, we can add 2π 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 3.

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. To be conservative, we also 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.

Clearly, our algorithm relies heavily on the compactness and convexity of the angle space Θ. 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 Θ allows us to conveniently refine the grid by separately cutting each coordinate of $\overline{\underline{\Theta}}^{(m)}$ into smaller pieces through simple division. In comparison, it is much harder to carry out this seemingly trivial arithmetic calculation in a clean manner with D-dimensional Euclidean coordinates on $\mathbb{S}^{D-1}$, due to the lack of convexity and the troublesome dependence across coordinates.

We find the current algorithm to be conservative and perform reasonably well in our simulation study and empirical illustration. In particular, see Figures 4 and 5 in Section 5.2 for more illustration of our computation algorithm. Admittedly, there are certainly some aspects of our algorithms that can be improved, especially in terms of its computational efficiency and grid uniformity under the great-circle metric. We look forward to future work

that improves the implementation of our computational procedure in the spherical-coordinate reparameterization.

5 Simulation

In this section, we examine the finite-sample performance of our estimation method via a Monte Carlo simulation study. We start by graphically presenting the estimated argmin set to illustrate the typical output of our method. Next, we study the performance of the first-stage nonparametric estimator $\hat{\gamma}$, or more precisely, the performance of the plugged-in estimator $G(\hat{\gamma})$, upon which our sample criterion function is built. Then, we show how the two-stage estimator $\hat{\beta}$ performs under various configurations of the data generating processes. Finally, we investigate how our proposed estimator performs under the lack of the point identification.

Adaptive-Grid Computation Algorithm

We first illustrate a typical output of our second-step computation algorithm based on the adaptive-grid search over the angle space, and show that the computation algorithm itself indeed works well. For this purpose we consider a simplified DGP, where there is no 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 idisyncratic shock to be $\epsilon_{ijt} \sim_{i.i.d.} TIEV$, so that we can thus calculate the true conditional choice probability conditioned on each $\mathbf{X}_i$. As in this section we are only seeking to illustrate the validity of the algorithm itself, we set N to be large with $N = 10^7$ and $D = 3, J = 3, T = 2$. We then apply our sadaptive-grid algorithm to search for $\overline{\beta}_0$.

Figure 4 shows how our computational algorithm works in finding the true unknown $\overline{\theta}_0$, the angle representation of the true $\overline{\beta}_0$ in the Θ space. In this figure the horizontal and the 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 $\widehat{Q}$. The black box indicates the area that the minimizers for the sample criterion $\widehat{Q}$ lie within, or more precisely, a rectangular enclosure of the estimated argmin set. The big black dot stands for the true parameter value $\overline{\theta}_0 = (0.4205, 0.4636)'$.

It is evident from Figure 4 that our adaptive-grid algorithm is able to correctly locate an area that covers the true $\overline{\theta}_0$, 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

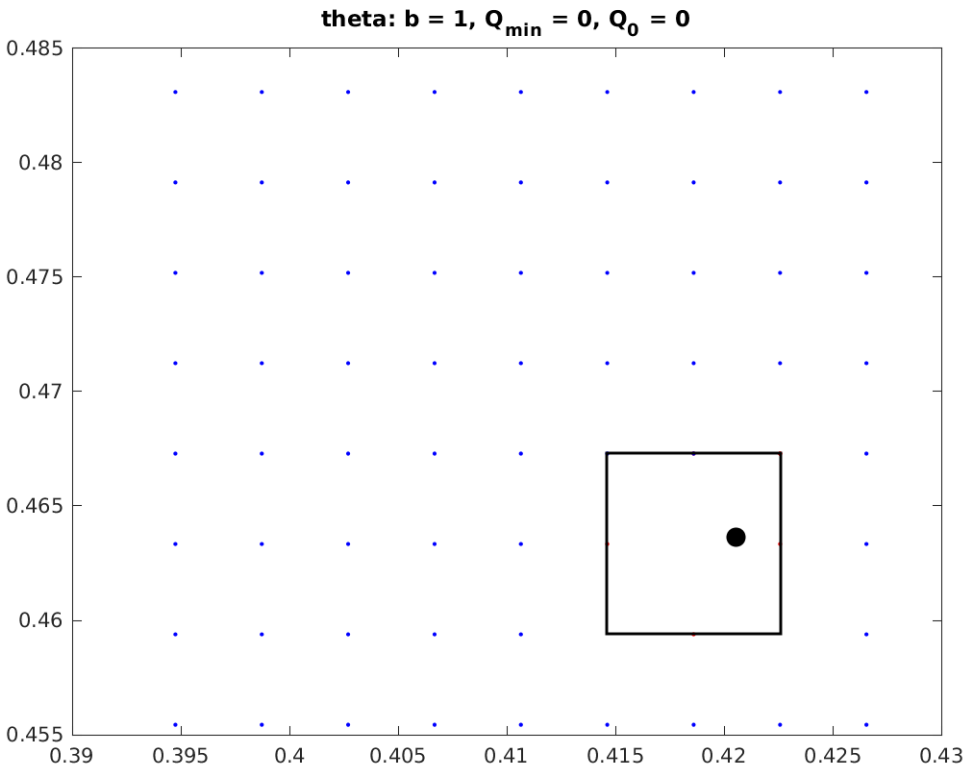


Figure 4: The Estimated Set for $\overline{\theta}_0$

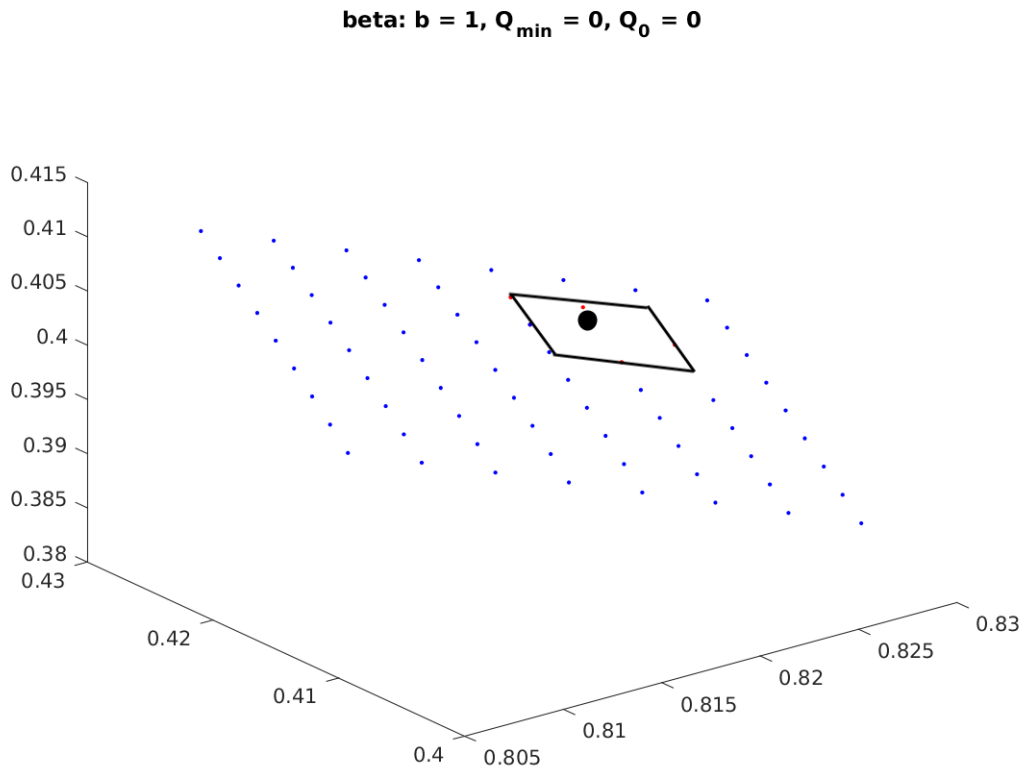


Figure 5: The Estimated Set for $\overline{\beta}_0$

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, our algorithm evaluates the criterion function $\widehat{Q}$ on a relatively small set of points shown by those blue and red dots in Figure 4 while achieving a desired level of accuracy.

For a more transparent representation of our estimates, we translate the angles θ in the polar coordinates into unit vectors β on the unit sphere $\mathbb{S}^2$, which is shown in Figure 5.

Figure 5 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 $\widehat{Q}$; the black box shows an enclosing set of the minimizers of $\widehat{Q}$. The big black dot represents the true parameter value $\overline{\beta}_0$, which resides inside the black box of the minimizers of $\hat{Q}$. It again illustrates that our computation algorithm is able to identify a tight area around the true parameter value $\overline{\beta}_0$.

Setup of Simulation Study

We have shown graphically that our adaptive-grid computation algorithm works reasonably well as as an optimization algorithm over the unit sphere under the angle-space reparameterization. Now, we describe the setup of our simulations study.

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

$$u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) = A_{ij}^{(0)}\left(X_{ijt}'\beta_0 + A_{ij}^{(1)}\right) + \epsilon_{ijt},$$

where $A_{ij}^{(0)}$ is an unobserved scale fixed effect that captures agent-level heteroskedasticity in utilities, and $A_{ij}^{(1)}$ is an unobserved location shifter specific to each agent-product pair. We fix $A_{ij}^{(0)}$ across j, and abbreviate it as A_{i0}. We further denote $A_{ij}^{(1)}$ as A_{ij} whenever no ambiguity is present. We draw the scale fixed effect A_{i0} from a uniform distribution on $[2, 2.5]$. We set the location fixed effect corresponding to product 1 to be zero, i.e., $A_{i1} \equiv 0$. To generate correlation between the observable characteristics $\mathbf{X}_i$ and the fixed effects $\mathbf{A}_i$, we introduce a latent variable $Z_i \sim_{i.i.d.} \mathcal{N}(0, 1)$ for each agent i. Then we construct the location fixed effect A_{i2} with respect to the product 2 to be the positive part of Z_i, i.e. $A_{i2} = [Z_i]_+$. The same latent variable Z_i will also be used to generate one covariate of the observable characteristics X_{ijt} later on. For all other products $j \in \{3, ..., J\}$, we draw the corresponding location shifter A_{ij} from a uniform distribution on $[-0.25, 0.25]$.

We draw the unobserved random utility shock ϵ_{ijt} independently from the Type I Extreme Value ($TIEV$) distribution with its location being 0 and scale of 1. Note that our estimation method does not require the knowledge of the distribution of ϵ_{ijt} per se. In the next section where we evaluate the performance of our first-step estimator, it will be used to calculate the true conditional choice probabilities. Specifically, the $TIEV$ assumption allows us to analytically compute the conditional choice probabilities given $\mathbf{X}_{it}$ and $\mathbf{A}_i$, and subsequently compute the true $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is})$ by numerically integrating over the conditional distribution of $\mathbf{A}_i$ by simulation. Note that the $TIEV$ assumption on ϵ makes the choice probabilities invariant under any common location shift in the indexes $X_{ijt}'\beta_0$.

We set the true $\beta_0 \in \mathbb{R}^D$ to be $(2, \underbrace{1, ..., 1}_{D-1})'$, and seek to estimate the direction of β_0, represented by the normalized vector $\overline{\beta}_0 := \beta_0 / \|\beta_0\|$ on the unit sphere $\mathbb{S}^{D-1}$. We will keep the specifications on $\underline{u, \mathbf{A}, \boldsymbol{\epsilon}, \beta_0}$ to be the same across all subsequent simulations.

In our *baseline* DGP configuration, we construct the $N \times D \times J \times T$ observable characteristic matrix $\mathbf{X}$ as follows. We draw $X_{ijt}^{(1)}$, the first coordinate of D coordinates

of X_{ijt}, from a uniform distribution on $[-1, 1]$. For the second coordinate, we set $X_{ijt}^{(2)} = W_{ijt} + Z_i$ with $W_{ijt} \sim_{i.i.d.} \mathcal{N}(0, 2J)$, inducing correlation and nonlinear dependence between A_{i2} and $X_{ijt}^{(2)}$ for each ijt tuple. For any pair of $(\mathbf{X}_{it}, \mathbf{X}_{is})$, the variable Z_i enters into the $2J$ random variables in $(X_{ijt})_{j=1}^{J}$ and $(X_{ijs})_{j=1}^{J}$. By configuring $Var(W_{ijt}) = 2J$, we control the variance of Z_i to be the same as the aggregate variance of all idiosyncratic components $(W_{ijt})_{j=1}^{J}$ and $(W_{ijs})_{j=1}^{J}$ from the two periods (t, s). More detailed discussion on this point will be provided below. For all other dimensions $d \in \{3, ..., D\}$, we draw $X_{ijt}^{(d)} \sim_{i.i.d.} \mathcal{N}(0, 1)$. Later on, we will modify these assumptions on $\mathbf{X}_{it}$ as well as vary (N, D, J, T) to evaluate the performance of our method under different scenarios.

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 based on the DGP in this section. Then we calculate the binary individual choice $\mathbf{Y}$ matrix according to model (1). Lastly, we compute our estimator $\hat{\beta}$ from the simulated observable data of $(\mathbf{X}, \mathbf{Y})$, and finally compare our estimator $\hat{\beta}$ with the true parameter value $\overline{\beta}_0$. Unless otherwise noted, all results are derived using the *baseline* DGP configuration. We will explicitly mentiond any alternations in the configurations of DGP relative to the baseline setup.

5.1 First-Stage Performance

We examine the performance of our first stage estimator $\hat{\gamma}$ or $G(\hat{\gamma})$ in this part. First, we calculate the true γ or $G(\gamma)$ using the knowledge of DGP. Specifically, for each of the $M = 100$ simulations, we numerically compute the true first-stage conditional expectation

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

for each (j, t, s) combination and for each realization of $(\mathbf{X}_{it}, \mathbf{X}_{is}) = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. We plug the true γ into the known function G to obtain the true $G(\gamma)$, which is used as the benchmark of comparison. Next, we estimate γ 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 exploit the $TIEV$ distribution of $\boldsymbol{\epsilon}$ to obtain the conditional choice probability according to the following formula:

$$\mathbb{E}\left[y_{ijt}-y_{ijs}|\,\mathbf{X}_{it}=\overline{\mathbf{X}},\mathbf{X}_{is}=\underline{\mathbf{X}},\mathbf{A_i}\right]=\frac{\exp^{A_{i0}\left(\overline{X}_j'\beta_0+A_{ij}\right)}}{\sum_{k=1}^{J}\exp^{A_{i0}\left(\overline{X}_k'\beta_0+A_{ik}\right)}}-\frac{\exp^{A_{i0}\left(\underline{X}_j'\beta_0+A_{ij}\right)}}{\sum_{k=1}^{J}\exp^{A_{i0}\left(\underline{X}_k'\beta_0+A_{ik}\right)}}.$$

Then we numerically integrate $\mathbf{A}_i$ out to obtain the true γ function

$$\begin{aligned}\mathbb{E}\left[y_{ijt}-y_{ijs}|\,\mathbf{X}_{it},\mathbf{X}_{is}\right]&=\int\mathbb{E}\left[y_{ijt}-y_{ijs}|\,\mathbf{X}_{it},\mathbf{X}_{is},\mathbf{A_i}\right]d\mathbb{P}_{t,s}\left(\mathbf{A_i}|\,\mathbf{X}_{it},\mathbf{X}_{is}\right)\\&=\int\left[\frac{\exp^{A_{i0}\left(X_{ijt}'\beta_0+A_{ij}\right)}}{\sum_{k=1}^{J}\exp^{A_{i0}\left(X_{ikt}'\beta_0+A_{ik}\right)}}-\frac{\exp^{A_{i0}\left(X_{ijs}'\beta_0+A_{ij}\right)}}{\sum_{k=1}^{J}\exp^{A_{i0}\left(X_{iks}'\beta_0+A_{ik}\right)}}\right]\\&\quad d\mathbb{P}_{t,s}\left(\mathbf{A_i}|\,\mathbf{X}_{it},\mathbf{X}_{is}\right)\end{aligned}\tag{27}$$

Because in our setting only A_{i2} is correlated with $\mathbf{X}$, we need the conditional probability distribution of $A_{i2}|\,\mathbf{X}_{it},\mathbf{X}_{is}$ to numerically evaluate the integral (27). Define $\overline{X}_{i,ts}^{(2)}:=\frac{1}{2J}\sum_{j=1}^{J}\left(X_{ijt}^{(2)}+X_{ijs}^{(2)}\right)\sim\mathcal{N}(0,1)+\mathcal{N}(0,1)$, then

$$A_{i2}|_{\mathbf{X}_{it},\mathbf{X}_{is}}\sim A_{i2}|_{\left(X_{ijt}^{(2)},X_{ijs}^{(2)}\right)_{j=1}^{3}}\sim A_{i2}|_{\overline{X}_{i,ts}^{(2)}}\sim\left[\mathcal{N}\left(\frac{1}{2}\overline{X}_{i,ts}^{(2)},\ \frac{1}{2}\right)\right]_{+}\tag{28}$$

We use simulation method to numerically calculate the integral (27). Specifically, we use $M_0:=10^7$ random draws of $\mathbf{A}_i\rightarrow\left\{\mathbf{A}_i^{(m)}:m=1,...,M\right\}$ according to the true conditional distribution $\mathbf{A}_i|_{(\mathbf{X}_{it},\mathbf{X}_{is})=\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)}$ in (28). For each j, we compute the numerical average

$$\frac{1}{M_0}\sum_{m=1}^{M_0}\left[\frac{\exp^{A_{i0}^{(m)}\left(\overline{X}_j'\beta_0+A_{ij}^{(m)}\right)}}{\sum_{j=1}^{J}\exp^{A_{i0}^{(m)}\left(\overline{X}_j'\beta_0+A_{ij}^{(m)}\right)}}-\frac{\exp^{A_{i0}^{(m)}\left(\underline{X}_j'\beta_0+A_{ij}^{(m)}\right)}}{\sum_{j=1}^{J}\exp^{A_{i0}^{(m)}\left(\underline{X}_j'\beta_0+A_{ij}^{(m)}\right)}}\right]$$

which we refer to as the true intertemporal differences in choice probabilities $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$ and enables us to derive the $G(\gamma)$ for any known G functional.

Next, we estimate $\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)$ by a machine learning algorithm with single-layered neural networks, using the `R` package "`mlr`" by Bischl et al. (2016). For each fixed product j, we regress $(y_{ijt}-y_{ijs})$ on neural-network functions of $vec\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)$. We tune over hyperparameters of the number of neurons, initial random weights and maximum number of iterations based on a three-fold cross validation. Then we use the tuned learner to obtain prediction $\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)$ of $\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)$ for each realized pair of $\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)$. Finally, we

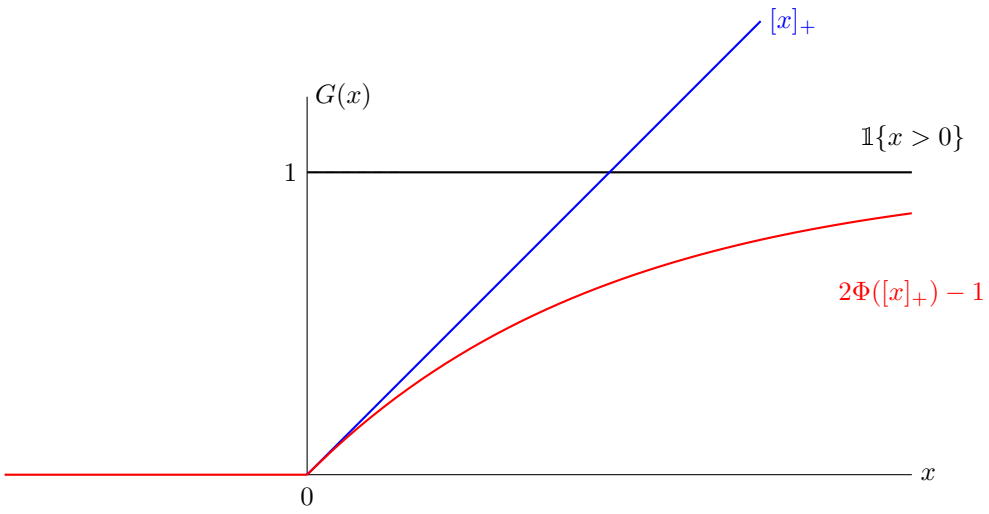


Figure 6: $G(\cdot)$ Functions

plug these predicted $\hat{\gamma}$'s into $G(\cdot)$ function and compare it to the true $G(\gamma)$ we obtained via simulation above.

Per the discussion in Section 3, our method only requires $G(\cdot)$ to be one-sided sign preserving, i.e., $G(z)$ is zero for nonpositive z and positive when for positive z. It is worth noting that the precision of our estimation of β_0 does not crucially depend on the accuracy of the first stage estimator $\hat{\gamma}$ per se, but only relies on how well $G(\hat{\gamma})$ approximates $G(\gamma)$. We compare below the performance of $G(\hat{\gamma})$ under several choices of G: the indicator function, the positive part function and an adjusted normal CDF $2\Phi\left([z]_+\right) - 1$.

We report in Table 1 both the means and the maximums of the mean squared errors across M simulations to evaluate the performance of our first stage estimator $G(\hat{\gamma})$. The first row of Table 1 lists the three choices of the one-sided sign preserving function G. Note that in the normal CDF $2\Phi\left([\hat{\gamma}]_+\right) - 1$ case, we first keep the positive part of $\hat{\gamma}$ obtained through the machine learning method, then evaluate it with normal CDF and renormalize the whole function to have a range between 0 and 1. 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 mean squared error of $G(\hat{\gamma})$ in the m^{th} simulation. Similarly the second row, "max MSE", reports the maximum MSE of $G(\hat{\gamma})$.

From Table 1 we see that the normal CDF $G(\cdot)$ performs the best in terms of both mean MSE and max MSE, while the indicator function gives the worst results and that the peformance of the positive part function lies somewhere in between. This is expected

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

because when the true γ is close to zero, it is more likely to have the estimated sign of $\hat{\gamma}$ to be different from γ. 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 γ is positive and large, it actually does not matter for our method whether the exact value of γ is estimated well by $\hat{\gamma}$. All we need is the sign of $\hat{\gamma}$ coincides with the sign of γ so as to obtain identifying restrictions on β_0. According to this intuition, the positive part function $[\hat{\gamma}]_+$ is expected to perform the least well when γ is positive and large. The adjusted normal CDF function $2\Phi\left([\hat{\gamma}]_+\right) - 1$ performs the best, as 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 γ_+ when γ is positive and large. For this reason, we will use the adjusted normal CDF function as G functional in our second-stage search for $\hat{\beta}$.

5.2 Two-Stage Performance

Using $G(\hat{\gamma}) = 2\Phi\left([\hat{\gamma}]_+\right) - 1$ in our finite sample criterion function $\hat{Q}(\theta)$ as defined in (19), we minimize $\hat{Q}(\theta)$ using our adaptive-grid algorithm in the angle space Θ. Later on we translate the estimated $\hat{\theta} \in \Theta$ back to $\hat{\beta} \in \mathbb{S}^{D-1}$ for transparency and compare it with the normalized $\overline{\beta}_0 \equiv \beta_0 / \|\beta_0\|$.

In this section, we first give a graphical illustration of the output of our algorithm from one of the M simulations under a regular setting. Then we show the simulation results under the baseline DGP configuration. Next, we study the performance of our algorithm under different numbers of individuals N, dimensions of observable characteristics D, numbers of products available J, and numbers of time periods T. Finally, we inspect how informative our estimates will be under the lack of point identification.

A Typical Set Estimator

In this section, we show graphically how a typical estimator $\hat{\beta}$ of $\overline{\beta}_0$ performs in one of the $M = 100$ simulations. The DGP remains the same as in the baseline configuration. We first estimate $\mathbb{E}[y_{ijt} - y_{ijs} | \mathbf{X}_{it}, \mathbf{X}_{is}]$, the intertemporal difference in conditional choice

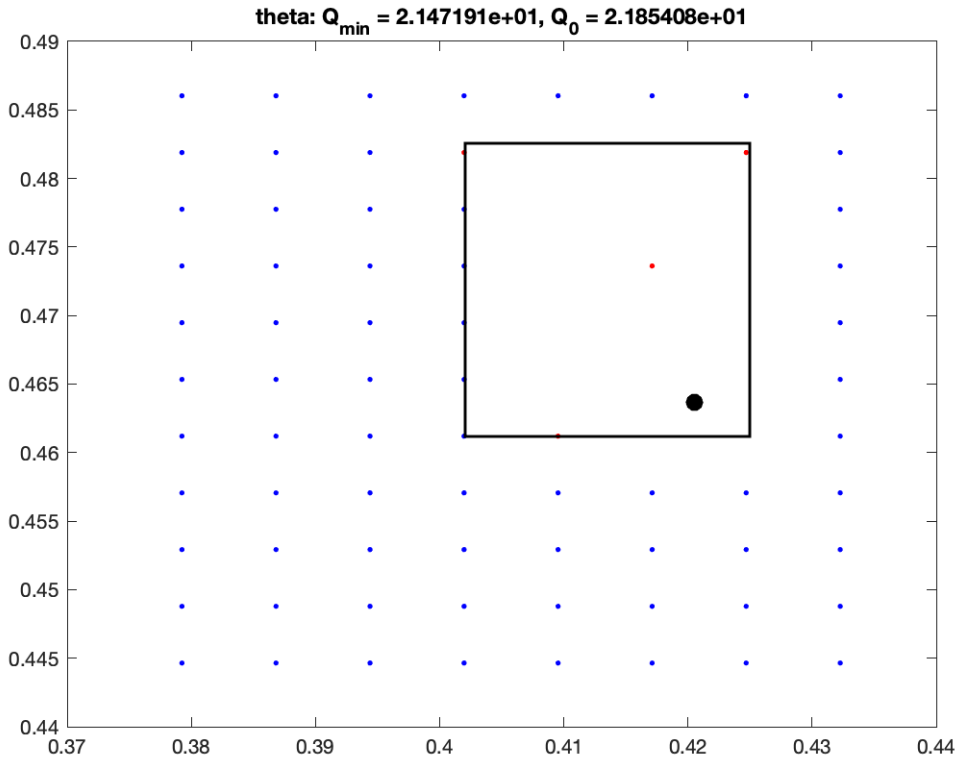


Figure 7: A Typical Estimator for $\overline{\theta}_0$

probabilities. Then we apply the second step adaptive-grid algorithm to search for the set of minimizers $\hat{\beta}$ of the sample criterion $\widehat{Q}$ on the unit sphere $\mathbb{S}^{d-1}$. We run the analysis under the setting of $N = 10,000, D = 3, J = 3, T = 4$. Below we show graphically a typical estimator $\hat{\theta}$ and its corresponding $\hat{\beta}$ from one of the M simulations.

The blue dots, black boxes and big black dots represent, respectively, the points that do not minimize the sample criterion, a rectangular enclosure of the points that minimize the criterion function, and the true parameter value $\overline{\theta}_0$ or $\overline{\beta}_0$ in Figure 7 and 8. We see that our method still produces a reasonably good estimator of $\overline{\theta}_0$ or $\overline{\beta}_0$, which lie within the black boxes. Moreover, the sizes of the black boxes are small in the absolute scale. Figure 7 and 8 graphically illustrate the typical outputs of our methods across the subsequent simulations. In the following, we use the center of these black boxes as the point estimator and report the summary statistics including the root MSE and the mean norm deviations (MND) across all M simulations based on our point estimator.

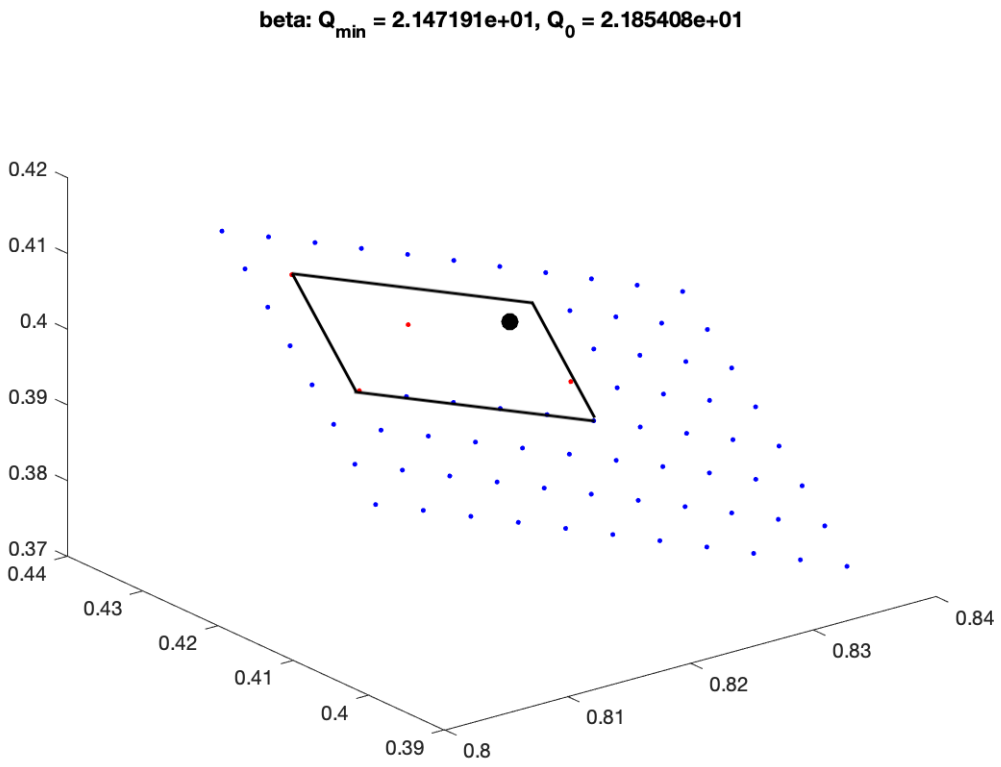


Figure 8: A Typical Estimator for $\overline{\beta}_0$

Table 2: Baseline Performance

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^m - \tilde{\beta}_{0,d}\right)$	-0.0050	0.0021	0.0006
upper bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^u - \tilde{\beta}_{0,d}\right)$	0.0015	0.0084	0.0108
lower bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^l - \tilde{\beta}_{0,d}\right)$	-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	$\sqrt{\frac{1}{M}\sum_m \left\|\hat{\beta}^m - \tilde{\beta}_0\right\|^2}$		0.0745	
mean norm deviations	$\frac{1}{M}\sum_m \left\|\hat{\beta}^m - \tilde{\beta}_0\right\|$		0.0648	

We now present some formal measures of the performance of our estimator under the baseline configuration.

Baseline Results

For the baseline configuration we set $N = 10,000, D = 3, J = 3, T = 2$. We define the set estimator as

$$\hat{B} := \arg\min_{\beta \in \mathbb{S}^{D-1}} \hat{Q}(\beta),$$

and for each dimension of product characteristics $d = 1, ..., D$, define

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

For each dimension $d = 1, ..., D$, $\hat{\beta}_d^u$ is defined as the maximum value along dimension d of the identified set $\hat{B}$, $\hat{\beta}_d^l$ as the minimum value along dimension d of the identified set $\hat{B}$, and $\hat{\beta}_d^m$ as the middle point along dimension d of the identified set $\hat{B}$. As

$$\hat{B} \subseteq \times_{d=1}^{D} \left[\hat{\beta}_d^l, \hat{\beta}_d^u\right],$$

we will refer to $\times_{d=1}^{D} \left[\hat{\beta}_d^l, \hat{\beta}_d^u\right]$ as the enclosing rectangle of $\hat{B}$. Note that if β_0 is point identified on $\mathbb{S}^{D-1}$, the enclosing rectangle $\times_{d=1}^{D} \left[\hat{\beta}_d^l, \hat{\beta}_d^u\right]$ should shrink to β_0 asymptotically.

Table 2 summarizes the main results for the simulations under our baseline configuration. 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 bias against the true $\overline{\beta}_0$ across all $M = 100$ simulations. The bias is very

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

small across all three dimensions with a magnitude between -0.0050 and 0.0021. The next two rows show the biases in estimating $\overline{\beta}_{0,d}$ using $\hat{\beta}_d^u$ and $\hat{\beta}_d^l$ respectively. Again, the bias is close to zero across all three dimensions. 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 $\tilde{\beta}_0$. In the second part of Table 2 we report the rMSE and MND of our second stage estimation results using $\hat{\beta}^m$. Our proposed algorithm is able to achieve a low rMSE and MND.

Results Varying N

In this section we vary N while holding $D = 3, J = 3, T = 2$ to show how our procedure performs under different size of observations. In addition to our baseline setup with $N = 10,000$, we calculate mean absolute deviation, 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. Mean absolute deviation 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 mean norm deviation being equal to 0.1076 and 0.1405, respectively. It is worthwhile mentioning that in this baseline configuration the total number of time periods is set to the minimum of $T = 2$. Our method can extract information from each of the $T(T-1)$ ordered pairs of time periods, which increases quadratically with 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 $10,000$. Table 4 summarizes the main results compared to the benchmark case $N = 1,000$.

Table 4 shows that our method achieves a convergence speed slower than the root-N rate. Compared with the case when $N = 1,000$, The relative ratios of rMSE are 1.68 for

Table 4: Convergence Rate

	$\sqrt{N/1000}$	$\text{rMSE}_{1000}/\text{rMSE}_N$	$\text{MND}_{1000}/\text{MND}_N$
$N = 10,000$	$\sqrt{10} \approx 3.2$	$\frac{0.1690}{0.0745} \approx 2.27$	$\frac{0.1405}{0.0648} \approx 2.17$
$N = 4,000$	$\sqrt{4} = 2$	$\frac{0.1690}{0.1006} \approx 1.68$	$\frac{0.1405}{0.0884} \approx 1.59$

Table 5: Performance Varying D, J, T

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

$N = 4,000$ and 2.27 for $N = 10,000$. In terms of MND, the corresponding ratios are 1.59 for $N = 4,000$ and 2.17 for $N = 10,000$. The results reported here suggests that our estimator may achieve a convergence rate faster than $N^{1/3}$.

Results Varying D, J, T

Now we fix $N = 10,000$ and vary D, J, T relative to the baseline configuration. Specifically, we draw A and X according to the following specifications:

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

which coincides with the baseline configuration at $D = 3, J = 3$.

We report in Table 5 the rMSE and the MND of our estimators for each of the corresponding configurations across all M simulations. First, as discussed earlier, a larger T improves the performance of our estimator, because we can now extract more information from $T \times (T-1)$ ordered pairs of time periods. Second, increase in D or J will adversely affect the performance of our estimator, but its magnitude 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. In summary, in datasets with larger T, which is more and more practically relevant with the increasing availability of long panel datasets, we find that our method produces good estimates in settings with higher dimensions of observable characteristics and larger choice sets of alternative products.

Estimation without Point Identification

We investigate in this section the performance of our proposed estimator under specifications where point identification fails. To make things comparable, we fix $N = 10,000, D = 3, J = 3, T = 2$ as in the baseline configuration, but we modify the baseline configuration in two different ways. We maintain the point identification of $\overline{\beta}_0$ in one setting but lose the point identification in the other setting.

For the setting where the point identification of $\overline{\beta}_0$ is preserved, we draw

$$Z_i \sim U\left[-\sqrt{3}, \sqrt{3}\right], \quad X_{ijt}^{(d)} \sim \begin{cases} U\left[-1,1\right], & d=1, \\ Z_i + \mathcal{N}\left(0,6\right), & d=2, \\ \mathcal{N}\left(0,1\right), & d=3, \end{cases} \tag{29}$$

and maintain the data generating assumptions on the fixed effects A_{ij}'s as in the baseline configuration. This DGP ensures point identification of $\overline{\beta}_0$ according to Appendix C. Note that the nonlinear structure in the indirect utility and the dependence between X and A are both preserved.

Next, we induce failure of point identification of $\overline{\beta}_0$ by discretizing and bounding the supports of the observable characteristics:

$$Z_i \sim U\left[-\sqrt{3}, \sqrt{3}\right], \quad X_{ijt}^{(d)} \sim \begin{cases} U\left\{-1,1\right\} & d=1, \\ Z_i + U\left[-\sqrt{6}, \sqrt{6}\right] & d=2, \\ U\left[-1,1\right] & d=3, \end{cases} \tag{30}$$

Specifically, from (30) to (29), we change the distribution of $X_{ijt}^{(1)}$ from $U\left[-1,1\right]$ to a Binomial distribution on two points $\{-1,1\}$ with equal probability. We also change the distributions of $X_{ijt}^{(2)}$ and $X_{ijt}^{(3)}$ to be uniformly distributed with same means, but set the boundaries of the supports to be one standard deviation from the mean, i.e, from $\mathcal{N}\left(0,6\right)$ to $U\left[-\sqrt{6}, \sqrt{6}\right]$ and from $\mathcal{N}\left(0,1\right)$ to $U\left[-1,1\right]$. Due to the discreteness and boundedness of $\mathbf{X}_i$ in (30), intertemporal differences in $X_{ijt} - X_{ijs}$ can no longer span all directions on the unit sphere, so the point identification fails in this case.

Note that we deliberately control the location and scale of each variable to be comparable across the two configurations (29) and (30), with the only differences being the presence of discreteness and boundedness. Table 6 contains simulation results under both configurations.

From Table 6, 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. Here all the

Table 6: Performance with and without Point ID

point ID ?	$\sum_d \text{mean(u-l)}_d$	$\sum_d \lvert\text{bias}_d\rvert$	rMSE	MND
(i) yes	0.0414	0.0119	0.0770	0.0661
(ii) no	0.0283	0.0185	0.0881	0.0762

Table 7: Performance with and without Point ID: Further Examination

point ID ?	rMSE			point ID ?	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	**(i)** yes	0.0661	0.0685	0.0697
(ii) no	0.0881	0.0892	0.0892	**(ii)** no	0.0762	0.0778	0.0778

results are calculated using $\hat{\beta}^m$, the middle point along each dimension of the estimated set $\hat{B}$. According to the second column of Table 6 there is minor change between the two settings in the average sizes of the estimated sets, both of which are relatively tight. In terms of mean absolute deviation, rMSE and MND, our estimator performs quite satisfactorily even under the lack of the point identification of $\overline{\beta}_0$.

In Table 7, we calculate the performance measures using the upper bound estimator $\hat{\beta}^u$ and the lower bound estimator $\hat{\beta}^l$ of the estimated set $\hat{B}$. When comparing the results between row (i) and (ii), the estimated sets tend to become larger and farther away from the true $\overline{\beta}_0$ when the point identification no longer holds. However, the changes are mild in magnitude, suggesting that our method handles discrete and bounded characteristics well, and remains informative without the point identification assumptions.

6 Empirical Illustration

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, permitting rich unobserved heterogeneity in factors such as brand loyalty or responsiveness to subtle flavor and packaging designs, which may affect choices in complex ways. The results show that our procedure produces estimates that conform well with economic intuition.

6.1 Data Description

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. It also includes additional information on store and product characteristics.

We choose to focus on the sales of popcorn among a huge variety of products covered by the Nielsen data. One of the reasons why we focus on popcorn is heuristically due to the consideration that purchases of popcorn are more likely to be driven by temporary urges of consumption without too much dynamic planning. Another reason lies in the observation that there is good variation in the display promotion status of popcorn. It is often easier to move popcorn to special display areas, say near the entrance or check-out alnes than some other products, such as milk and eggs, which need to be refrigerated at all time. The variation in display promotion status 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 defined by Nielsen for year 2015. There are $T = 52$ weeks in 2015, giving us $T_0 = T \times (T-1) = 2652$ ordered week pairs. The Nielsen data contains detailed universal product code (UPC) level information, which we use to aggregate the data into brand level observations. We focus on the top 3 brands ranked by market share and aggregate the rest into a fourth product - "all other products". There is a "fifth brand" that represents the "outside option".

The Nielsen Scanner Data contains data on the volume sold by each UPC re-scaled to the same units. To calculate the market share variable, we aggregate the volume vairable across all the UPCs under brand j in DMA i during week t and divide it by the total volume sold for all brands in the same DMA and week. We will use it as the dependent variable for the first-stage nonparametric regression.

The observed product characteristics $\mathbf{X}$ for each brand include price, promotion status and their interaction term. We calculate brand j's price in DMA i in week t Price_{ijt} as the weighted average unit price equaling to the total weekly sales of all the UPCs contained in brand j in DMA i during week t divided by total selling volume of the same UPCs in the same DMA and week. In the Nielsen data we find two variables related to promotion: display and feature. The feature variable captures whether or not a product is promoted via advertisement on local newspapers, free standing inserts, free standing circulars or online from the retailer's website. The display variable reflects whether or not a product is brought temporarily to the store lobby, front of store or end of aisle to increase its exposure. Due to their similarity, we define the promotion variable Promo_{ijt} to be the maximum of feature and

Table 8: Empirical Application: Summary Statistics

	mean	s.d.	min	max
DMA-level Market Share s_{ijt}	25.00%	21.59%	0.07%	96.69%
Price_{ijt}	0.4924	0.1803	0.1094	1.3587
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

display and calculate it as the percentage of stores in DMA i in week t that either feature or display brand j, i.e. $\text{Promo}_{ijt} = (\text{feature} \vee \text{display})_{ijt}$. The third variable we construct is the interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$. It is included to show the effect of promotion of a product on the price elasticity of consumers. The summary statistics of the variables discussed above are provided in Table 8.

6.2 Methodology

We use the observed DMA-level market shares as an estimate of

$$s_{ijt} = \mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i \right].$$

Then to apply our method, under the strong stationarity assumption we still need to run the first-stage estimation of

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

Specifically, we nonparametrically regress $(s_{ijt} - s_{ijs})$ on $(\mathbf{X}_{it}, \mathbf{X}_{is})$ through the single-layered neural network from the `mlr` package in `R`. We obtain an estimator $\hat{\gamma}$ of the true $\gamma_{i,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[s_{ijt} - s_{ijs} \middle| \left(\mathbf{X}_{it}, \mathbf{X}_{is}\right) = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \right]$ evaluated at each realized pair of $\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right) = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ in the data. Then we make use of the first-stage estimators to run our adaptive-grid search algorithm, obtaining an estimator $\hat{\theta}$ that minimizes sample criterion function $\hat{Q}\left(\theta\right)$ on the corresponding spherical coordinate space. Finally, we translate $\hat{\theta}$ back onto the unit sphere of $\mathbb{R}^3$, as a unit vector $\hat{\beta}$, for easier interpretation of our results.

Table 9: Empirical Application: Estimation Results

	$\hat{\beta}^{mean}$	$\left[\hat{\beta}^{l}, \hat{\beta}^{u}\right]$
Price_{ijt}	-0.9681	[-0.9687, -0.9677]
Promo_{ijt}	0.1988	[0.1861, 0.2078]
$\text{Price}_{ijt} \times \text{Promo}_{ijt}$	0.1520	[0.1399, 0.1700]

Table 10: Empirical Application: Comparison of Results

	$\hat{\beta}^{mean}$	$\hat{\beta}^{OLS}$	$\hat{\beta}^{OLS-FE}$	$\hat{\beta}^{MLogit-FE}$
Price_{ijt}	-0.9681	0.0240	-0.3803	-0.8511
Promo_{ijt}	0.1988	0.5760	0.5978	0.4589
$\text{Price}_{ijt} \times \text{Promo}_{ijt}$	0.1520	-0.8171	-0.7057	-0.2552

6.3 Results and Discussion

We report our estimation results for this empirical application in Table 9. The column corresponding to $\hat{\beta}^{mean}$ in Table 9 uses the center of the identified set $\hat{B}$ as the estimated coefficients for each variable of $\mathbf{X}$. $\left[\hat{\beta}^{l}, \hat{\beta}^{u}\right]$ corresponds to the lower and upper bound of the identified set $\hat{B}$. Our method estimate the coefficient for Price_{ijt} to be negative in the range of [-0.9687, -0.9677]. This is in line with the economic theory of a downward sloping demand curve. The estimated sign for the coefficient on Promo_{ijt} is positive with a magnitude between 0.1861 and 0.2078. It gives direct evidence that featuring a product in advertisements or displaying it in the prominent location of a store will help increase sales of the product.

The most interesting result is the estimated coefficient on the interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$, which falls in $[0.1399, 0.1700]$, a strictly positive interval. An intuitive explanation for the positive sign we found 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 people become less price-sensitive for these specially promoted products.

To further illustrate the advantages of our method, we compare $\hat{\beta}^{m}$with the estimates obtained through three other different popular methods, i.e. OLS, OLS with scalar-valued fixed effects and the multinomial logit with fixed effects. Results are summarized in the Table 10. The estimated coefficients under different methods are each normalized to have

norm 1. The OLS regression result shows that the estimated coefficient on Price_{ijt} is 0.0240, a slightly positive number, while the estimated coefficient on the interaction term is -0.8171, a large negative number. We find these results counterintuitive and unreasonable. The results obtained by OLS with fixed effects are reported in the third column of Table 10, which performs slightly better than OLS, but the estimated coefficient for the interaction term is stil negative. Similar results are also obtained under a multinomial logit model with fixed effects, which again is unable to generate a positive coefficient on the interaction term, $\text{Price}_{ijt} \times \text{Promo}_{ijt}$. We regard the sharp contrast between our result and the results obtained in these alternative methods as an empirical illustration that by accomodating 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 procudues economically more reasonable results than the aforementioned alternative methods.

7 Extension and Generalization

7.1 Counterfactual Analysis

So far we have focused on the identification and estimation of the index parameter β_0. While β_0 may be the only parameter of interest in many settings, often times we are also interested in counterfactual parameters defined as some functional of not only β_0 but also other unknown components of the model. In this extension, we discuss how the estimate $\hat{\beta}$ of β_0, and the computed indexes based on $\hat{\beta}$, may be used to estimate more sophisticated counterfactual parameters.

An important class of counterfactual parameters concerns with the prediction of counterfactual market shares (aggregate choice probability), say, in the form of

$$\mu\left(\overline{\mathbf{X}}\right) := \int \mathbb{E}\left[y_{ijt} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}_i, \mathbf{A}_i\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i\right).$$

In the context of our empirical illustration with popcorn sales (or more generally, the retailing industry), a marketing campaign manager might be interested in predicting the effects of a specific promotion strategy, potentially via a combination of price discounts and in-store special displays, and thus optimizes over promotion strategies. Moreover, demand elasticities may be further computed as $\nabla\mu\left(\overline{\mathbf{X}}\right)$, which gives the marginal effect of an exogenous change in certain observable characteristics on consumer choices.

It is important to note that in the expression of μ we use the marginal distribution $\mathbb{P}\left(\mathbf{A}_i\right)$ rather than the conditional distribution $\mathbb{P}\left(\mathbf{A}_i | \, X_{it} = \overline{\mathbf{X}}_i\right)$. This separation between

the exogenously imposed counterfactual $\overline{\mathbf{X}}$ and the distribution of the unobserved $\mathbf{A}_i$ is key to the interpretation of $\mu\left(\overline{\mathbf{X}}\right)$ as the direct effect of the exogenous change in observable characteristics $\mathbf{X}$ on choice probabilities, with the unobserved heterogeneity $\mathbf{A}$ unaffected by this exogenous change held *fixed*.

To achieve this separation, we seek to identify and estimate the integrand $\mathbb{E}\left[y_{ijt}|\,\mathbf{X}_{it}=\overline{\mathbf{X}}_i,\mathbf{A}_i\right]$, which is a function of β_0. We first define the scalar index values

$$\delta_{ijt} := X_{ijt}^{'}\beta_0.$$

Let $\overline{\boldsymbol{\delta}} = \left(\overline{\delta}_{i1t},...,\overline{\delta}_{iJt}\right)^{'} = \overline{\mathbf{X}}_i^{'}\beta_0$. Conditional on $\boldsymbol{\delta}_{it}=\overline{\boldsymbol{\delta}}$ and $\mathbf{A}_i$, by our model specification we have

$$\mathbb{E}\left[y_{ijt}|\,\boldsymbol{\delta}_{it}=\overline{\boldsymbol{\delta}},\mathbf{A}_i\right] = \psi_j\left(\overline{\boldsymbol{\delta}},\ \mathbf{A}_i\right) =: \psi_{ij}\left(\overline{\boldsymbol{\delta}}\right).$$

Here an important observation is that although the individual heterogeneity $\mathbf{A}_i$ is not directly observable, the identity of i is observable. We can hold individual i fixed in the regression to control for $\mathbf{A}_i$ and only use variations in the data across t in a long panel setting to estimate the conditional choice probability. Specifically, suppose we have long panels, i.e. $T\to\infty$. We can identify $\psi_{ij}\left(\overline{\boldsymbol{\delta}}\right)$ by

$$\psi_{ij}\left(\overline{\boldsymbol{\delta}}\right) = \mathbb{E}\left[y_{ijt}|\,\boldsymbol{\delta}_{it}=\overline{\boldsymbol{\delta}},\quad i\right]$$

and estimate it via nonparametric regression of y_{ijt} on $(\delta_{i1t},...,\delta_{iJt})$ for each fixed i across $t=1,...,T$. We can use kernel, sieve, local linear regression or neural network techniques to derive the nonparametric estimator $\hat{\psi}_{ij}$ for ψ_{ij}.

We are now in the position to evaluate the counterfactual market share of product j at any counterfactual $\overline{\mathbf{X}}_i$. We first use the estimated $\hat{\beta}$ to compute the counterfactual index $\hat{\boldsymbol{\delta}}$ evaluated at $\overline{\mathbf{X}}_i$

$$\hat{\boldsymbol{\delta}}\left(\overline{\mathbf{X}}_i\right) = \left(\hat{\delta}_{i1t},...,\hat{\delta}_{iJt}\right)^{'} = \overline{\mathbf{X}}_i\hat{\beta}.$$

Then we obtain $\hat{\psi}_{ij}\left(\hat{\boldsymbol{\delta}}\right)$ by plugging $\hat{\boldsymbol{\delta}}$ into the nonparametric estimate $\hat{\psi}_{ij}$ for each fixed i. Finally, we can obtain an estimate of μ by averaging over individuals in the sample:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{ij}\left(\hat{\boldsymbol{\delta}}_i\left(\overline{\mathbf{X}}_i\right)\right).$$

Notice that the parametric index structure $\overline{\boldsymbol{\delta}} = \overline{\mathbf{X}}_i\beta_0$ provides the key foundation for counterfactual extrapolation to some $\overline{\mathbf{X}}_i$ that may not lie in the support of the (in-sample) observed characteristics $\mathbf{X}_i$, provided that that the support of the in-sample *indexes* $\mathbf{X}_i\beta_0$ is large enough.

The validity of such counterfactual analysis relies on the functional form assumption of the index structure, which provides the underlying global invariance property for extrapolation. Yet without any assumption that induces some form of global invariance structures, it is conceptually impossible to make "out-of-support" counterfactual extrapolations. For example, if our linear index $\overline{\boldsymbol{\delta}} = \overline{\mathbf{X}}_i\beta_0$ is replaced by a fully nonparametric function $\overline{\boldsymbol{\delta}} = f\left(\overline{\mathbf{X}}_i\right)$, then it is no longer feasible to carry out "out-of-support" counterfactual prediction as described above, even if we have estimated f well on its on the support of $\mathbf{X}_i$.

Admittedly there are

7.2 Monotone Multi-Index Models

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

$$\gamma\left(\mathbf{X}_i\right) = \phi\left(\mathbf{X}_i\beta_0\right) \tag{31}$$

where:

- $(y_i, \mathbf{X}_i)_{i=1}^N$ constitutes a random sample of N observations on a scalar[8] random variable y_i and a $J \times D$ random matrix $\mathbf{X}_i$.
- $\gamma\left(\overline{\mathbf{X}}\right) = \mathcal{T}\left(F_{y_i|\mathbf{X}_i=\overline{\mathbf{X}}}(\cdot)\right)$ 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 (31) 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 unknown finite-dimensional parameter.[9] Again, we normalize β_0, as $\beta_0 \in \mathbb{S}^{D-1}$, as β_0 is at best identified up to scale given that ϕ is an unknown function.

As in Lee (1995), Powell and Ruud (2008) and Ahn, Ichimura, Powell, and Ruud (2018), model (31) restricts the dependence of $\gamma\left(\mathbf{X}_i\right)$ on the matrix $\mathbf{X}_i$ to the J linear parametric

[8] Similar to Ahn, Ichimura, Powell, and Ruud (2018), the dimension of y_i is largely irrelevant to the analysis of model (31): it is the dimension of γ that matters. Nevertheless, for the clarity of presentation, we take y_i to a scalar variable.

[9] If $\beta_0 = \mathbf{0}$, then model (31) degenerates to $\gamma\left(\mathbf{X}_i\right) \equiv \phi\left(\mathbf{0}\right)$, which is a constant. As $\gamma\left(\mathbf{X}_i\right)$ is assumed to be an identifiable quantity, whether $\beta_0 = \mathbf{0}$ or not is also identifiable. We focus thereafter on the nondegenerate case where $\beta_0 \neq \mathbf{0}$.

indexes $\mathbf{X}_i\beta_0 \equiv \left(X_{ij}'\beta_0\right)_{j=1}^{J}$.[10]

A noteworthy difference of model (31) from the setup in AIPR is that we take $\gamma(\mathbf{X}_i)$ here both to be scalar-valued quantities, while AIPR require their $\gamma(\mathbf{X}_i)$ to have dimension, namely R, no lower than J. This "order condition" $R \geq J$ is necessary for their vector-valued version of function ϕ to admit a left-inverse ϕ^{-1} such that $\phi^{-1}(\gamma(\mathbf{X}_i)) = \mathbf{X}_i\beta_0$, which constitutes the foundation for their subsequent analysis. However, we impose no such order condition for the sake of invertibility, as we will not rely on invertibility at all. Instead, we adopt a multinomial version of the monotonicity assumption that is familiar in the literature on monotone single-index models, such as Han (1987), Ahn, Ichimura, and Powell (1996), Cavanagh and Sherman (1998), and Abrevaya (2000).

Assumption 7 (Weak Monotonicity)**.** *ϕ 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 ϕ, model (31) builds in the fundamental lack of additive separability across the parametric indexes. As demonstrated later in Section 2, the key idea developed below for the general multi-index model (31) naturally apply to the analysis of the panel multinomial choice model under complete lack of additive separability.

However, we now provide a few illustrative examples for model (31) that satisfy Assumption 7 beyond multinomial choice settings.

Example 3 (Sample Selection Model)**.** Consider a sample selection model studied by Heckman (1979):

$$\begin{aligned} y_i^* &= W_i'\mu_0 + u_i \\ d_i &= \mathbb{1}\left\{Z_i'\lambda_0 + v_i \geq 0\right\} \\ y_i &= y_i^* \cdot d_i \end{aligned}$$

[10]Note that model (31) is without loss of generality relative to the following seemingly more general formulation, in which β_0 is explicitly allowed to be heterogeneous across the J rows of $\mathbf{X}_i$:

$$\gamma(\mathbf{X}_i) = \phi\left(\left(X_{ij}'\beta_{0j}\right)_{j=1}^{J}\right),$$

where X_{ij}, β_{0j} are both vectors of dimension d_{x_j} for each $j \in \{1, ..., J\}$, making the unknown parameter of interest $\beta_0 := \left(\beta_{01}', ..., \beta_{0J}'\right)'$ a $\sum_{j=1}^{J} D_j$-dimensional vector. This, however, could be readily incorporated in model (31) by appropriately redefining $\tilde{\mathbf{X}}_i$ so that each $\tilde{X}_{ij}$ is a vector of $\sum_{j=1}^{J} D_j$-dimensional vector with nonzero entries (given by X_{ij}) only at the D_j corresponding positions, giving the representation $\gamma\left(\tilde{\mathbf{X}}_i\right) = \phi\left(\tilde{\mathbf{X}}_i\beta_0\right)$ as in model (31).

where y_i^* is the latent variable that is observable only when $d_i = 1$. 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 positive correlation. Then we have

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

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 (31) with Assumption 7 satisfied.

Example 4 (Dyadic Network Formation Model under Nontransferable Utilities)**.** Consider the following simple dyadic network formation model under nontransferable utilities (NTU):

$$D_{ij} = \mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ij}^{'}\gamma_0 \geq \epsilon_{ij}\right\}\mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ji}^{'}\gamma_0 \geq \epsilon_{ji}\right\}, \tag{32}$$

where $W_{ij} \equiv W_{ji}$ denotes some symmetric observable characteristics between a pair of individuals ij, while (Z_{ij}, Z_{ji}) denote some asymmetric observable characteristics between a pair of individuals 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 indicator 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 (31) 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 (32). 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, 2018), 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.

The next proposition generalizes our key identification result (Theorem 1) to the setting of monotone multi-index models:

Proposition 3 (General Identifying Restriction)**.** *Under model* (31) *with Assumption 7, for any* $\overline{\mathbf{X}}, \underline{\mathbf{X}} \in Supp\left(\mathbf{X}_i\right)$*,*

$$\gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right) \;\Rightarrow\; \text{NOT}\;\left(\left(\overline{X}_j - \underline{X}_j\right)\beta_0 \leq 0,\;\forall j = 1, ..., J\right). \tag{33}$$

Notice that Proposition 3 applies to all functionals γ 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 γ. 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 γ into a joint restriction on β_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. Section 7.2 provides a discussion about monotone multi-index models, and Gao, Li, and Xu (2018) considers a dyadic network formation model under nontransferable utilities, where the key method proposed in this paper can be applied.

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.

Finally, on the technical side, the spherical-coordinate reparameterization adopted in this paper is shown to enjoy several theoretical and practical niceties in estimation and computation, which may also be useful in a larger class of semiparametric discrete-response models without scale identification. Meanwhile, asymptotic distributions and inferential procedures for our estimators, developed with the nonstandard formulation of moment conditions, the spherical geometry induced by the lack of scale identification, and the built-in discreteness of the Boolean algebra in mind, are also conceptually important, technically interesting and practically relevant research questions beyond the scope of the panel multinomial choice setting considered in this paper. Some of these questions are currently under investigation by the authors in separate projects, while a broader range of aspects are left for future research.

References

Abrevaya, J. (2000): "Rank estimation of a generalized fixed-effects regression model," *Journal of Econometrics*, 95, 1–23.

Ahn, H., H. Ichimura, and J. L. Powell (1996): "Simple estimators for monotone index models," Manuscript, Department of Economics, UC Berkeley.

Ahn, H., H. Ichimura, J. L. Powell, and P. A. Ruud (2018): "Simple estimators for invertible index models," *Journal of Business & Economic Statistics*, 36.

Aradillas-Lopez, A. (2018): "A Comment on "Simple Estimators for Invertible Index Models"," *Journal of Business & Economic Statistics*, 36, 18–21.

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 (1995a): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

——— (1995b): "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.

Cavanagh, C. and R. P. Sherman (1998): "Rank estimators for monotonic index models," *Journal of Econometrics*, 84, 351–381.

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. Newey (2017): "Nonseparable multinomial choice models in cross-section and panel data," *arXiv preprint arXiv:1706.08418*.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models 1," *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 (2018): "Logical Differencing in Network Formation Models under Nontransferable Utilities," 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 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*, New York: Wiley.

Khan, S., F. Ouyang, and E. Tamer (2017): "Adaptive inference in 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.

Luarn, P. and H.-H. Lin (2003): "A customer loyalty model for e-service context." *J. Electron. Commerce Res.*, 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.

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," Tech. rep., 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.

Wasserman, L. (2013): *All of statistics: a concise course in statistical inference*, Springer Science & Business Media.

Appendix

A Proof of Theorem 2

Lemma 3. *The reparameterized population criterion function $Q : \Theta \to \mathbb{R}_+$ is continuous and*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}\sum_{j}\sum_{t\neq s} G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\theta\right) - Q\left(\theta\right)\right| \xrightarrow{p} 0. \tag{34}$$

Proof. Recall that

$$\begin{aligned} Q_{j,t,s}\left(\theta\right) &= \mathbb{E}\left[G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\theta\right)\right] \\ &= \int G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\theta\right)\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right), \end{aligned}$$

and thus

$$\begin{aligned} &\left|Q_{j,t,s}\left(\overline{\theta}\right) - Q_{j,t,s}\left(\underline{\theta}\right)\right| \\ \leq &\int G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\left|\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\overline{\theta}\right) - \lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\underline{\theta}\right)\right|\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right) \\ = &\int G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\mathbb{1}\left\{\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\overline{\theta}\right) \neq \lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\underline{\theta}\right)\right\}\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right) \\ = &\mathbb{E}\left[G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\mathbb{1}\left\{\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\overline{\theta}\right) \neq \lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\underline{\theta}\right)\right\}\right]. \end{aligned} \tag{35}$$

Notice that

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

which is continuous in each $\overline{\theta}\in\Theta$ and $\underline{\theta}\in\Theta$ with probability one, since $v_k\left(\mathbf{X}_{it}-\mathbf{X}_{is}\right)$ has no mass point for each (k,t,s).

Then, as also $(\mathbf{X}_{it},\mathbf{X}_{is})$ is i.i.d. across i, Θ is compact, and the indicator function is bounded, all conditions for Lemma 2.4 in Newey and McFadden (1994) are satisfied, by which we deduce that the expectation term in (35) is continuous in $\overline{\theta}$ and $\underline{\theta}$. Consequently, we have

$$\left|Q_{j,t,s}\left(\overline{\theta}\right) - Q_{j,t,s}\left(\underline{\theta}\right)\right| \leq \mathbb{E}\left[\mathbb{1}\left\{\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\overline{\theta}\right) \neq \lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\underline{\theta}\right)\right\}\right] \to 0,$$

as $\rho_\Theta\left(\overline{\theta},\underline{\theta}\right)\to 0$, giving the continuity of $Q=\sum_j\sum_{t\neq s}Q_{j,t,s}$ on Θ.

Moreover, by Newey and McFadden (1994, Lemma 2.4) we also have

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\theta\right)-Q_{j,t,s}\left(\theta\right)\right|\xrightarrow{p}0,$$

which implies (34) after summation over (j,t,s). □

Lemma 4. *Under Assumptions 2, 5 and 6, we have*

$$\sup_{\theta\in\Theta}\left|\hat{Q}\left(\theta\right)-Q\left(\theta\right)\right|=O_p\left(c_N\right).$$

Proof. By the Lipschitz continuity of G in Assumption 5, we have

$$\begin{aligned}
&\left|\hat{Q}_{j,t,s}\left(\theta\right)-Q_{j,t,s}\left(\theta\right)\right|\\
\leq&\frac{1}{N}\sum_i\left|G\left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)-G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right)\right|\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\theta\right)\\
&+\left|\frac{1}{N}\sum_i\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};\theta\right)-\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};\theta\right)\right]\right)\right|\\
\leq&\frac{1}{N}\sum_i c\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|+O_p\left(N^{-\frac{1}{2}}\right)\\
=&c\int\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)+O_p\left(N^{-\frac{1}{2}}\right)\\
&+c\frac{1}{N}\sum_i\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|-c\int\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\\
=&c\int\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)+O_p\left(N^{-\frac{1}{2}}\right)+O_p\left(N^{-\frac{1}{2}}\right)\\
\leq&c\sqrt{\int\left|\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)-\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)\right|^2\mathrm{d}\mathbb{P}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right)}+O_p\left(N^{-\frac{1}{2}}\right)\\
=&O_p\left(c_N\right)+O_p\left(N^{-\frac{1}{2}}\right)\\
=&O_p\left(c_N\right)
\end{aligned}$$

and hence we have

$$\sup_{\theta\in\Theta}\left|\hat{Q}\left(\theta\right)-Q\left(\theta\right)\right|=O_p\left(c_N\right).$$

□

Main Proof of Theorem 2

Proof. We verify Condition C.1 in Chernozhukov, Hong, and Tamer (2007) so as 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 is satisfied 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. Then, Theorem 3.1.(1) in Chernozhukov, Hong, and Tamer (2007) implies the consistency of $\hat{\theta}$. □

B Pairwise Time Homogeneity of Errors

As mentioned in Section 2.2, Assumption 3 is stronger than necessary, and our identification strategy (and Proposition 1) carries over under the weaker Assumption 3', which requires that

$$\boldsymbol{\epsilon}_{it} | \left(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i \right) \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i \right)$$

To see why Proposition 1 continue to hold, notice that

$$\begin{aligned}
\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) &= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \middle| \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{A}_i\right] - \mathbb{E}\left[y_{ijs} | \, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \middle| \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \middle| \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right]
\end{aligned}$$

with

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

where $\tilde{\boldsymbol{\epsilon}}_i$ denotes generic realizations of $\boldsymbol{\epsilon}_{it}$ and $\boldsymbol{\epsilon}_{is}$ conditional on $\mathbf{X}_{it}, \mathbf{X}_{is}$ and $\mathbf{A}_i$:

$$\tilde{\boldsymbol{\epsilon}}_i \sim \boldsymbol{\epsilon}_{it} \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right).$$

Writing $\overline{\boldsymbol{\delta}} = \overline{\mathbf{X}}\beta_0$ and $\underline{\boldsymbol{\delta}} = \underline{\mathbf{X}}\beta_0$. If

$$\overline{\delta}_j \leq \underline{\delta}_j \text{ and } \overline{\delta}_k \geq \underline{\delta}_k \text{ for all } k \neq j,$$

we have

$$\mathbb{1}\left\{u\left(\delta_{ijt}, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\delta_{ikt},\, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \leq \int \mathbb{1}\left\{u\left(\delta_{ijs}, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\delta_{iks},\, A_{ik}, \tilde{\epsilon}_{ik}\right)\right\}$$

for all realizations of $\mathbf{A}_i$ and $\tilde{\boldsymbol{\epsilon}}_i$, so that

$$\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \leq 0$$

for all realizations of $\mathbf{A}_i$, which further implies that

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \leq 0.$$

Taking the logical contraposition again gives Proposition 1.

C Consistency under Point Identification

In this section, we prove consistency results under sufficient conditions for the point identification of β_0. For simplicity of notation, we fix $T = 2$ and denote $\Delta X_{ij} = X_{ij1} - X_{ij2}$ for all individual i and product j. We use the indicator function in this section for $G(\cdot)$. First we list two additional assumptions on the support of $\Delta\mathbf{X}_i$, either of which by itself is sufficient for the point identification of β_0.

Assumption 8 (Continuous Support of $\Delta\mathbf{X}_i$)**.** *There exists some $\epsilon > 0$ such that $B_\epsilon(0) \subseteq Supp\left(\Delta X_{ij} \middle| \Delta X_{il},\, l \neq j\right)$ for all $j \in \{1, ..., J\}$.*

Assumption 9 (Discrete Support of $\Delta\mathbf{X}_i$)**.** *For some $k \in \{1, ..., d_x\}$ that satisfies $\beta_0^k \neq 0$. $Supp\left(\Delta X_{ij}^k \middle| \Delta X_{il},\, l \neq j\right) = \mathbb{R}$ for all $j \in \{1, ..., J\}$, and 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 8 is satisfied when (X_{ij}) is continuous random vector. On the other hand, Assumption 9 can accommodate discrete regressors generally, but requires one continuous covariate with large support. Assumption 8 or 9 on $Supp\left(\Delta X_{ij}\right)$ ensures that following

inequalities hold simultaneously with strictly positive probability

$$\begin{cases} \Delta X_{ij}^{'}\beta_0 > 0 \\ \Delta X_{ik}^{'}\beta_0 < 0 \quad \forall k \neq j. \end{cases}$$

Given the above assumptions, we may without loss of generality normalize

$$\|\beta_0\| = 1$$

and consider in the parameter space

$$\mathbb{B} := \left\{ \beta \in \mathbb{R}^{d_x} : \|\beta\| = 1 \right\}.$$

Next, we define

$$\begin{aligned} Q_j^+(\beta) &= \mathbb{E}\left[\tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta)\right] \\ \tau_j^+\left(\overline{\mathbf{X}}\right) &= \mathbb{1}\left\{\mathbb{E}\left[\Delta y_{ij} | \mathbf{X}_i = \overline{\mathbf{X}}\right] > 0\right\} \\ \lambda_j^+\left(\overline{\mathbf{X}};\beta\right) &= \mathbb{1}\left\{\left(\Delta \overline{X}_j^{'}\beta \leq 0\right)\right\} \cdot \prod_{k\neq j} \mathbb{1}\left\{\Delta \overline{X}_k^{'}\beta \geq 0\right\} \end{aligned}$$

and similarly for $Q_j^-(\beta)$, $\tau_j^-(\mathbf{X}_i)$ and $\lambda_j^-(\mathbf{X}_i;\beta)$. Construct the population criterion function Q as

$$\begin{aligned} Q(\beta) &= \mathbb{E}\left[\sum_j \left(\tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta) + \tau_j^-(\mathbf{X}_i)\,\lambda_j^-(\mathbf{X}_i;\beta)\right)\right] \\ &= \sum_j \mathbb{E}\left[\tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta)\right] + \sum_j \mathbb{E}\left[\tau_j^-(\mathbf{X}_i)\,\lambda_j^-(\mathbf{X}_i;\beta)\right] \\ &= \sum_j Q_j^+(\beta) + \sum_j Q_j^-(\beta) \end{aligned}$$

and let $\widehat{Q}_n(\beta)$ be the sample analogue of $Q(\beta)$. Define $\hat{\beta}$ to be the minimizer of $\widehat{Q}_n(\beta)$ over $\mathbb{B}$. Next theorem states the consistency result for $\hat{\beta}$ under sufficient conditions for the point identification of β_0.

Theorem 3 (PMC Point Consistency)**.** *Under Assumption 1, 2, 3, and either 8 or 9, we have for any $\epsilon > 0$ and any $j \in \{0, 1, ..., J\}$, there exist $\delta > 0$ such that*

$$\inf_{\beta \in \mathbb{B}\setminus B_\epsilon(\beta_0)} Q(\beta) \geq Q(\beta_0) + \delta$$

Furthermore, if we also have Assumption 4,

$$\hat{\beta} \xrightarrow{p} \beta_0 \text{ as } n \to \infty.$$

Proof. To begin with, we prove the first part of Theorem 3, the point identification of β_0. We show the identification result using $Q_j^+(\beta)$ and break the argument into 5 steps. Then the point identification result follows immediately by the symmetry between $Q_j^+(\beta)$ and $Q_j^-(\beta)$ and a triangular inequality argument.

We first show that β_0 is a minimizer of $Q_j^+(\beta)$ for any fixed $j = j_0$.

From equation 6 we know for all $\mathbf{X}_i \in Supp(\mathbf{X}_i)$, $\tau_j^+(\mathbf{X}_i) = 1$ implies $\lambda_j^+(\mathbf{X}_i; \beta_0) = 0$, which means the integrand in $Q_j^+(\beta)$ achieves its minimum 0 at β_0. Therefore, β_0 is a minimizer of $Q_j^+(\beta)$.

Second, we show $\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i) = 1\right\} > 0$.

Recall that $\Delta X_{ij} := X_{ij1} - X_{ij2}$. Note by definition

$$\left\{\tau_j^+(\mathbf{X}_i) = 1\right\} \Leftrightarrow \left\{\phi\left(X_{ij1}'\beta_0, \left(-X_{ik1}'\beta_0\right)_{k\neq j}\right) > \phi\left(X_{ij2}'\beta_0, \left(-X_{ik2}'\beta_0\right)_{k\neq j}\right)\right\},$$

we have

$$\begin{aligned}
&\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i) = 1\right\} \\
=&\mathbb{P}\left\{\phi\left(X_{ij1}'\beta_0, \left(-X_{ik1}'\beta_0\right)_{k\neq j}\right) > \phi\left(X_{ij2}'\beta_0, \left(-X_{ik2}'\beta_0\right)_{k\neq j}\right)\right\} \\
\geq&\mathbb{P}\left\{\left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\} \\
>&0
\end{aligned}$$

where the last inequality by Assumption 8 or 9 and the first inequality by

$$\begin{aligned}
&\left\{\left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\} \\
\Rightarrow&\left\{\phi\left(X_{ij1}'\beta_0, \left(-X_{ik1}'\beta_0\right)_{k\neq j}\right) > \phi\left(X_{ij2}'\beta_0, \left(-X_{ik2}'\beta_0\right)_{k\neq j}\right)\right\}
\end{aligned}$$

due to Assumption 1.

Third, we show for $\forall \beta \in \mathbb{B}, \beta \neq \beta_0$

$$\mathbb{P}\left\{\left(\Delta X_{ij}'\beta \leq 0\right) \wedge \left(\Delta X_{ik}'\beta \geq 0, \forall k \neq j\right) \middle| \left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\} > 0.$$

Fix an arbitrary $\beta \neq \beta_0$. Denote the spatial angle between β and β_0 to be $\theta > 0$. Define

$$H_j := \{\mathbf{V} \in supp\,(\Delta \mathbf{X}_i) : \langle V_j, \beta\rangle \leq 0 < \langle V_j, \beta_0\rangle, \langle V_k, \beta_0\rangle < 0 \leq \langle V_k, \beta\rangle, \forall k \neq j\}$$

By continuity of inner product operator and Assumption 8 or 9, we know H_j has strict positive probability measure implied by $\theta > 0$. Therefore, by definition of conditional probability we have

$$\begin{aligned}
&\mathbb{P}\left\{\left(\Delta X_{ij}'\beta \leq 0\right) \wedge \left(\Delta X_{ik}'\beta \geq 0, \forall k \neq j\right) \middle| \left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\} \\
=&\frac{\mathbb{P}\left\{\Delta \mathbf{X}_i \in H_j\right\}}{\mathbb{P}\left\{\left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\}} \\
>&0
\end{aligned}$$

Fourth, we show the point-wise result that β_0 uniquely minimizes $Q_j^+(\beta)$. Note that

$$\begin{aligned}
Q_j^+(\beta) =&\mathbb{E}\left[\tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta)\right] \\
=&\mathbb{E}\left[\lambda_j^+(\mathbf{X}_i;\beta)\middle|\tau_j^+(\mathbf{X}_i)=1\right]\cdot\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i)=1\right\} \\
=&\mathbb{E}\left[\mathbb{1}\left\{\left(\Delta X_{ij}'\beta \leq 0\right)\right\}\cdot\prod_{k\neq j}\mathbb{1}\left\{\left(\Delta X_{ik}'\beta \geq 0\right)\right\}\middle|\tau_j^+(\mathbf{X}_i)=1\right]\cdot\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i)=1\right\} \\
=&\mathbb{P}\left\{\left(\Delta X_{ij}'\beta \leq 0\right) \wedge \left(\Delta X_{ik}'\beta \geq 0, \forall k \neq j\right)\middle|\tau_j^+(\mathbf{X}_i)=1\right\}\cdot\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i)=1\right\}
\end{aligned}$$

Because $\left\{\left(\Delta X_{ij}'\beta_0 > 0\right) \wedge \left(\Delta X_{ik}'\beta_0 < 0, \forall k \neq j\right)\right\}$ is a sufficient condition for $\left\{\tau_j^+(\mathbf{X}_i)=1\right\}$ and both events have strictly positive probability measure, we have for any $\beta \neq \beta_0$

$$\mathbb{P}\left\{\left(\Delta X_{ij}'\beta \leq 0\right) \wedge \left(\Delta X_{ik}'\beta \geq 0, \forall k \neq j\right)\middle|\tau_j^+(\mathbf{X}_i)=1\right\} > 0$$

from the result in step 3.

By Proposition 1 we have

$$\mathbb{P}\left\{\left(\Delta X_{ij}'\beta_0 \leq 0\right) \wedge \left(\Delta X_{ik}'\beta_0 \geq 0, \forall k \neq j\right)\middle|\tau_j^+(\mathbf{X}_i)=1\right\} = 0$$

Combining above two equations we derive for any $\beta \neq \beta_0$

$$\begin{aligned}
Q_j^+(\beta) =&\mathbb{P}\left\{\left(\Delta X_{ij}'\beta \leq 0\right) \wedge \left(\Delta X_{ik}'\beta \geq 0, \forall k \neq j\right)\middle|\tau_j^+(\mathbf{X}_i)=1\right\}\cdot\mathbb{P}\left\{\tau_j^+(\mathbf{X}_i)=1\right\} \\
>&0 \times \mathbb{P}\left\{\tau_j^+(\mathbf{X}_i)=1\right\} \\
=&\mathbb{P}\left\{\left(\Delta X_{ij}'\beta_0 \leq 0\right) \wedge \left(\Delta X_{ik}'\beta_0 \geq 0, \forall k \neq j\right)\middle|\tau_j^+(\mathbf{X}_i)=1\right\}\cdot\mathbb{P}\left\{\tau_j^+(X_i)=1\right\}
\end{aligned}$$

$$=Q_j^+(\beta_0)$$

So far, we have obtained the point-wise result that β_0 uniquely minimizes $Q_j^+(\beta)$ on $\mathbb{B}$. To obtain the point identification result, we need to show it uniformly minimizes $Q_j^+(\beta)$ on $\mathbb{B}$. Here we follow Newey and McFadden (1994) and show

(i) $Q_j^+(\beta)$ is continuous on $\mathbb{B}$, and
(ii) $\mathbb{B}$ is compact.

Note (ii) is automatically satisfied by construction of $\mathbb{B}$. For (i), we follow Lemma 2.4 of Newey and McFadden (1994) and show

(i.1) $g(\mathbf{X}_i;\beta) := \tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta)$ is continuous at each $\beta \in \mathbb{B}$ with probability one, and
(i.2) $\mathbb{E}\sup_{\beta\in\mathbb{B}}|g(\mathbf{X}_i;\beta)| < \infty$.

Here (i.1) is satisfied by continuity of inner product and Assumption 8 or 9. (i.2) is satisfied by the fact that $|g(\mathbf{X}_i;\beta)| \leq 1$ by construction.

We have proved that for any $\epsilon > 0$ and any $j \in \{0, 1, ..., J\}$, there exist $\delta > 0$ such that

$$\inf_{\beta\in\mathbb{B}\setminus B_\epsilon(\beta_0)} Q_j^+(\beta) \geq Q_j^+(\beta_0) + \delta.$$

The point identification result in the first part of Theorem 3 follows immediately from the symmetry between $Q_j^+(\beta)$ and $Q_j^-(\beta)$ and a triangular inequality argument.

Next, we prove the latter part of Theorem 3, the consistency result of $\hat{\beta}$ for β_0, under the additional Assumption 4.

Following Newey and McFadden (1994) we need show the uniform convergence (UC) of $\widehat{Q}_n(\beta)$ to $Q(\beta)$,

$$\sup_{\beta\in\mathbb{B}}\left|Q(\beta) - \widehat{Q}_n(\beta)\right| \xrightarrow{p} 0 \quad \text{as } n \to \infty.$$

To prove UC, first we construct an infeasible estimator $\overline{\beta}$ of β_0 assuming we can observe the true τ^+ and τ^-

$$\overline{\beta} := \arg\max_{\beta\in\mathbb{B}} \overline{Q}_n(\beta)$$

where

$$\overline{Q}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_j \left(\tau_j^+(\mathbf{X}_i)\,\lambda_j^+(\mathbf{X}_i;\beta) + \tau_j^-(\mathbf{X}_i)\,\lambda_j^-(\mathbf{X}_i;\beta)\right)\right]$$

Then we can see conditions for Lemma 2.4 of Newey and McFadden (1994) are all satisfied, thus the uniform convergence of $\overline{Q}_n(\beta)$ to $Q(\beta)$ is established

$$\sup_{\beta\in\mathbb{B}}\left|Q(\beta) - \overline{Q}_n(\beta)\right| \xrightarrow{p} 0 \quad \text{as } n\to\infty$$

Second, observe that

$$\begin{aligned}
&\sup_{\beta\in\mathbb{B}}\left|\widehat{Q}_n(\beta) - \overline{Q}_n(\beta)\right| \\
=&\sup_{\beta\in\mathbb{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[\lambda_j^+(\mathbf{X}_i;\beta)\left(\hat{\tau}_j^+(\mathbf{X}_i) - \tau_j^+(\mathbf{X}_i)\right) + \lambda_j^-(\mathbf{X}_i;\beta)\left(\hat{\tau}_j^-(\mathbf{X}_i) - \tau_j^-(\mathbf{X}_i)\right)\right]\right| \\
\leq&\sup_{\beta\in\mathbb{B}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[\left|\lambda_j^+(\mathbf{X}_i,\beta)\right|\left|\left(\hat{\tau}_j^+(\mathbf{X}_i) - \tau_j^+(\mathbf{X}_i)\right)\right| + \left|\lambda_j^-(\mathbf{X}_i;\beta)\right|\left|\left(\hat{\tau}_j^-(\mathbf{X}_i) - \tau_j^-(\mathbf{X}_i)\right)\right|\right] \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[\left|\left(\hat{\tau}_j^+(\mathbf{X}_i) - \tau_j^+(\mathbf{X}_i)\right)\right| + \left|\left(\hat{\tau}_j^-(\mathbf{X}_i) - \tau_j^-(\mathbf{X}_i)\right)\right|\right] \\
\leq&\sum_{j=0}^{J}\left\{\sup_{i=1,\ldots,n}\left|\hat{\tau}_j^+(\mathbf{X}_i) - \tau_j^+(\mathbf{X}_i)\right| + \sup_{i=1,\ldots,n}\left|\hat{\tau}_j^-(\mathbf{X}_i) - \tau_j^-(\mathbf{X}_i)\right|\right\} \\
\xrightarrow{p}&\ 0 \quad \text{as } n\to\infty
\end{aligned}$$

where the first inequality by triangular inequality, the second inequality by both $\left|\lambda_j^+(\mathbf{X}_i;\beta)\right|$ and $\left|\lambda_j^-(\mathbf{X}_i;\beta)\right|$ are bounded from above by 1 for all $\beta\in\mathbb{B}$ by construction, and the last convergence result by Assumption 4.

Therefore, we can see the UC condition of $\widehat{Q}_n(\beta)$ to $Q(\beta)$ holds following an triangular inequality argument

$$\begin{aligned}
\sup_{\beta\in\mathbb{B}}\left|Q(\beta) - \widehat{Q}_n(\beta)\right| &\leq \sup_{\beta\in\mathbb{B}}\left|Q(\beta) - \overline{Q}_n(\beta)\right| + \sup_{\beta\in\mathbb{B}}\left|\widehat{Q}_n(\beta) - \overline{Q}_n(\beta)\right| \\
&\xrightarrow{p} 0 \quad \text{as } n\to\infty
\end{aligned}$$

Now we are in the position to prove the consistency result. For any $\epsilon > 0$, there exist $\delta > 0$ such that

$$\begin{aligned}
\mathbb{P}\left\{\left\|\hat{\beta} - \beta_0\right\| \geq \epsilon\right\} =&\mathbb{P}\left\{\hat{\beta}\in\mathbb{B}\backslash B_\epsilon(\beta_0)\right\} \\
\leq&\mathbb{P}\left\{Q\left(\hat{\beta}\right) \leq Q(\beta_0) - \delta\right\}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{P}\left\{Q\left(\hat{\beta}\right)-\widehat{Q}_n\left(\hat{\beta}\right)+\widehat{Q}_n\left(\hat{\beta}\right)-\widehat{Q}_n\left(\beta_0\right)+\widehat{Q}_n\left(\beta_0\right)-Q\left(\beta_0\right)\leq-\delta\right\}\\
\leq&\mathbb{P}\left\{Q\left(\hat{\beta}\right)-\widehat{Q}_n\left(\hat{\beta}\right)+\widehat{Q}_n\left(\beta_0\right)-Q\left(\beta_0\right)\leq-\delta\right\}\\
\leq&\mathbb{P}\left\{2\sup_{\beta\in\mathbb{B}}\left|Q\left(\beta\right)-\widehat{Q}_n\left(\beta\right)\right|\geq\delta\right\}\\
\to&0\ \text{ as } n\to\infty
\end{aligned}
$$

where the first inequality by the identification result in the first part of Theorem 3, the second inequality by the definition of $\hat{\beta}$, and the last convergence result by the UC condition of $\widehat{Q}_n\left(\beta\right)$ to $Q\left(\beta\right)$. □

D More on Monotone Multi-Index Models

We now provide some further discussion on the monotone multi-index model (31) presented in Section (7.2), and explain the similarities with and differences from the methods proposed for *monotone single-index* models and *invertible multi-index* models.

In the extreme case with $J=1$ (and we write $\overline{\mathbf{X}}=\overline{X}$ to emphasize this degeneration), our multi-index setting essentially degenerates to the single-index setting, as studied by Manski (1987).

$$
\gamma\left(\overline{X}\right)>\gamma\left(\underline{X}\right)\Rightarrow\neg\left(\left(\overline{X}-\underline{X}\right)\beta\leq0\right)\Leftrightarrow\left(\overline{X}-\underline{X}\right)\beta>0 \tag{36}
$$

In this case, the method of maximum score or rank-order estimators pioneered by Manski (1987) would be applicable, due to a peculiar feature of the single-index setting that is not generalizable to our multi-index setting:

$$
\begin{cases}
\gamma\left(\overline{X}\right)>\gamma\left(\underline{X}\right)\Rightarrow\left(\overline{X}-\underline{X}\right)\beta>0\Rightarrow\gamma\left(\overline{X}\right)\geq\gamma\left(\underline{X}\right)\\
\gamma\left(\overline{X}\right)<\gamma\left(\underline{X}\right)\Rightarrow\left(\overline{X}-\underline{X}\right)\beta<0\Rightarrow\gamma\left(\overline{X}\right)\leq\gamma\left(\underline{X}\right)
\end{cases}
$$

which essentially encodes, and can often be strengthened to, the following *equivalence* relationship under suitable conditions:[11]

$$
\gamma\left(\overline{X}\right)>\gamma\left(\underline{X}\right)\Leftrightarrow\left(\overline{X}-\underline{X}\right)\beta>0, \tag{37}
$$

potentially with a probabilistic qualifier "almost surely".[12]. Consider taking $\gamma\left(\overline{X}\right)=\mathbb{E}\left[y_i|\,X_i=\overline{X}\right]$ where y_i is a *binary* variable as in the discrete choice models studied in

[11] Or a stronger form: $\gamma\left(\overline{X}\right)>\gamma\left(\underline{X}\right)\Leftrightarrow\left(\overline{X}-\underline{X}\right)\beta>0$.

[12] See, for example, Lemma 1 in Manski (1987), equations (6) and (7) in Abrevaya (2000)

these papers, the "equivalence-type" relationship (37) allows the formulation of the well-known maximum score or rank-order estimators based on maximizing the sample analogue of following population criterion function, a la Manski (1987)[13]:

$$\begin{aligned} S(\beta) &= \mathbb{E}\left[(y_i - y_j)\,\mathbb{1}\left\{(X_i - X_j)\,\beta > 0\right\}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[y_i - y_j \middle| X_i, X_j\right]\mathbb{1}\left\{(X_i - X_j)\,\beta > 0\right\}\right] \\ &= \mathbb{E}\left[\left[\gamma(X_i) - \gamma(X_j)\right]\mathbb{1}\left\{(X_i - X_j)\,\beta > 0\right\}\right] \end{aligned} \tag{38}$$

The proofs of identification (and consistency) based on this formulation, say, in Manski (1985, 1987), Abrevaya (2000) and Khan, Ouyang, and Tamer (2017), basically rely on the equivalence relationship (37). In our multi-index setup, however, an equivalence relationship in the form of (37) no longer holds in general:

$$\begin{aligned} \gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right) &\Rightarrow \neg\left(\left(\overline{X}_j - \underline{X}_j\right)\beta_0 \leq 0,\ \forall j = 1, ..., J\right) \not\Rightarrow \gamma\left(\overline{\mathbf{X}}\right) \geq \gamma\left(\underline{\mathbf{X}}\right), \\ \gamma\left(\overline{\mathbf{X}}\right) < \gamma\left(\underline{\mathbf{X}}\right) &\Rightarrow \neg\left(\left(\overline{X}_j - \underline{X}_j\right)\beta_0 \geq 0,\ \forall j = 1, ..., J\right) \not\Rightarrow \gamma\left(\overline{\mathbf{X}}\right) \leq \gamma\left(\underline{\mathbf{X}}\right), \end{aligned}$$

so the usual identification and estimation strategy based in single-index settings are no longer directly applicable due to the multi-index nature of the problem. [14] Our proposed solution is to instead exploit the much weaker yet more robust implication relationship (33).

Relatedly, in another extreme case where $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ are such that $\overline{X}_j = \underline{X}_j$ for all except one $j \in \{1, ..., J\}$, say, $\overline{X}_1 \neq \underline{X}_1$, then the identifying restriction (33) reduces to the following simpler form:

$$\gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right) \Rightarrow \neg\left(\left(\overline{X}_1 - \underline{X}_1\right)\beta \leq 0\right) \Leftrightarrow \left(\overline{X}_1 - \underline{X}_1\right)\beta > 0 \tag{39}$$

The similarity between (36) and (39) is not mere coincidence: by conditioning on the event that $\overline{X}_j = \underline{X}_j$ for all except one $j \in \{1, ..., J\}$, the multi-index setting is effectively reduced

[13] There are more general and flexible formulations of "scores" as well as smoothed versions of the sample criterion function (potentially with another form of scale normalization) as have been studied by, for example, Manski (1987), Han (1987), Horowitz (1992) and Abrevaya (2000). These various formulations involve subtle differences across them in terms of the exact sequential placement of conditional expectation operators, indicator (or sign) function, and minus signs (differencing operations), leading to technical differences in the proofs of identification and consistency. However, the key methodology underlying them is basically the same as represented in (38).

[14] In another logical direction, such "equivalence-type" relationship do not hold either.

$$\begin{aligned} \left(\overline{X}_j - \underline{X}_j\right)\beta_0 > 0,\ \forall j = 1, ..., J &\Rightarrow \gamma\left(\overline{\mathbf{X}}\right) \geq \gamma\left(\underline{\mathbf{X}}\right) \not\Rightarrow \left(\overline{X}_j - \underline{X}_j\right)\beta_0 \geq 0,\ \forall j = 1, ..., J, \\ \left(\overline{X}_j - \underline{X}_j\right)\beta_0 < 0,\ \forall j = 1, ..., J &\Rightarrow \gamma\left(\overline{\mathbf{X}}\right) \leq \gamma\left(\underline{\mathbf{X}}\right) \not\Rightarrow \left(\overline{X}_j - \underline{X}_j\right)\beta_0 \leq 0,\ \forall j = 1, ..., J. \end{aligned}$$

to a single-index setting. Khan, Ouyang, and Tamer (2017) explicitly focus on exploiting events of such types where all but one row of the observable characteristics $\overline{\mathbf{X}}$ and $\underline{\mathbf{X}}$ match exactly (i.e., all but one product's observable characteristics remain unchanged over time in their panel multinomial choice setting), and construct a rank-order estimator that exploits the equivalence relationship in the form of (37)[15] Shi, Shum, and Song (2018) partially exploit the same type of events (in their definition of G_I so as to reduce a summation over all products to a single term) in their identification analysis. However, a theoretical concern about this approach, which is also empirical to some extent, is that $\left\{\overline{X}_j = \underline{X}_j\right\}$ often regarded as probability-zero events if X_{ij} is continuously distributed. Hence, it may require more structures and assumptions (say, linearity, additivity, and continuity) to ensure these "probability-zero" events have bites. Our key idea in (33), on other hand, is fundamentally formulated based on inequalities only. The advantage of our approach is to explicitly take advantage of events defined by inequalities (heuristically "positive-probability events" if the distribution of, say, X_{ij} is absolutely continuous with respect to the Lebesgue measure), while automatically incorporating the equality-defined events such as $\left\{\overline{X}_j = \underline{X}_j\right\}$.

Now we discuss in more detail the differences between our approach and that developed in Ahn, Ichimura, Powell, and Ruud (2018, AIPR). First, as pointed out earlier, we do not impose any "order condition" on the dimensionality of γ that is necessary for invertibility, which we do not rely on. Hence, instead of working with a vector-valued γ function, our formulation features a scalar-valued function γ, or a real functional, of the conditional distribution of y_i given $\mathbf{X}_i$. Interpreted from another perspective, our approach applies separately to each element of the vector-valued function γ in AIPR, provided that our Assumption 7 is imposed. Second, as a consequence of the first to some extent, a more salient difference of our monotonicity-based approach to the invertibility approach adopted in in AIPR is that we are able to utilize all *inequality* relationships among scalar-valued γ, while AIPR relies on *equality* relationships among whole vectors of γ. As pointed out by the comment of Aradillas-Lopez (2018) on AIPR, invertibility, or the ability to "asymptotically match" pairs of observations with equal $\gamma(\mathbf{X}_i) = \gamma(\mathbf{X}_j)$ as in equations (15) and (16) of AIPR, may not be possible in a variety of microeconometric models. We emphasize that, when confronted with a vector-valued $\gamma = (\gamma_m)_{m=1}^M$, our method can be applied to all conditional events of the form $\{\gamma_m(\mathbf{X}_i) > \gamma_m(\mathbf{X}_h)\}$ separately across m (so the inequalities need not be aligned across m).

Finally, some may argue that our current approach discards information from potential equalities of (our scalar-valued) γ, which may also contributes to the identification and

[15] See the first displayed statement on page 6 in Khan, Ouyang, and Tamer (2017).

estimation of β_0. However, our method can be adapted to also incorporate information from equality-type events of the form $\left\{\gamma\left(\overline{\mathbf{X}}\right) = \gamma\left(\underline{\mathbf{X}}\right)\right\}$, provided that we strengthen Assumption 7 to a stronger version:

Assumption 10 (Strict Monotonicity)**.** *ϕ is nondegenerate and strictly increasing in each of its J arguments on Supp $(\mathbf{X}_i\beta_0) \subseteq \mathbb{R}^J$.*

Assumption 10 is often imposed in the literature, say, on panel multinomial choice models, via a more primitive assumption that the distribution of structural errors is absolutely continuous with respect to the Lebesgue measure.

Whenever Assumption 10 is imposed, we may derive an additional identifying restriction by exploiting the implication of an observed equality $\gamma\left(\overline{\mathbf{X}}\right) = \gamma\left(\underline{\mathbf{X}}\right)$:

Proposition 4 (Additional Identifying Restriction)**.** *Under model (31) with Assumption 10, for any $\overline{\mathbf{X}}, \underline{\mathbf{X}} \in Supp\left(\mathbf{X}_i\right)$,*

$$\gamma\left(\overline{\mathbf{X}}\right) = \gamma\left(\underline{\mathbf{X}}\right) \Rightarrow \text{NOT}\left[\begin{array}{c} \left(\overline{X}_j - \underline{X}_j\right)\beta_0 < 0,\ \forall j \text{ s.t. } \overline{X}_j \neq \underline{X}_j \\ \vee \left(\overline{X}_j - \underline{X}_j\right)\beta_0 > 0,\ \forall j \text{ s.t. } \overline{X}_j \neq \underline{X}_j \end{array}\right], \tag{40}$$

Again, compared to AIPR, Proposition 4 allows us to exploit equalities (or "matching") between scalars in the form of $\gamma\left(\overline{\mathbf{X}}\right) = \gamma\left(\underline{\mathbf{X}}\right)$, while AIPR requires matching of whole vectors of γ.

Under Assumption 10, our method essentially allows us to "exploit all the data": given a scalar-valued γ and any two $\overline{\mathbf{X}} \neq \underline{\mathbf{X}}$, either $\gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right)$, $\gamma\left(\overline{\mathbf{X}}\right) < \gamma\left(\underline{\mathbf{X}}\right)$ or $\gamma\left(\overline{\mathbf{X}}\right) = \gamma\left(\underline{\mathbf{X}}\right)$, in whichever case we may derive some potentially informative identifying restriction on the parameter of interest β_0. However, for the simplicity and clarity of presentation, we refrain from explicitly exploiting (40) in our subsequent analysis, with the understanding that (40) can be easily incorporated by an additional term in our construction of criterion functions.