Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects*†

Wayne Yuan Gao‡ and Ming Li§

January 6, 2026

Abstract

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

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

†We thank Xiaohong Chen, Peter Phillips, and Phil Haile for their invaluable advice and encouragement. We thank three anonymous referees for their comments that significantly improved the paper. We thank Don Andrews, Isaiah Andrews, Tim Armstrong, Xu Cheng, Tim Christensen, Ben Connault, Francis Diebold, Bo Honoré, Joel Horowitz, Yuichi Kitamura, Patrick Kline, Lixiong Li, Yuan Liao, Charles Manski, Aviv Nevo, Matt Seo, Xiaoxia Shi, Frank Schorfheide, Elie Tamer, Ed Vytlacil, Rui Wang, Sheng Xu and participants at various seminars and conferences for helpful comments. We thank Wenli Lyu and Chuyue Tian for excellent research assistance.

‡Gao: Department of Economics, University of Pennsylvania, waynegao@upenn.edu.

§Li: Department of Economics and Risk Management Institute, National University of Singapore, mli@nus.edu.sg.

1 Introduction

This paper proposes a method for semiparametric identification and estimation in panel multinomial choice models, where we allow for infinite-dimensional fixed effects that enter into consumer utilities in an additively nonseparable manner. The proposed method also applies more widely beyond panel multinomial choice models, and can be adapted to a wide range of models characterized by *multi-index single-crossing conditions*, which we introduce later in this paper.

To fix ideas, we start with the following panel multinomial choice model:

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

where agent i's utility from a candidate product j at time t, represented by $u(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt})$, 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 that can be heterogeneous across each agent-product combination. We emphasize that X_{ijt} and A_{ij} can be arbitrarily dependent. The last term ϵ_{ijt} is an idiosyncratic time-varying error term of arbitrary dimension. 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 $u(X_{ijt}'\beta_0, A_{ij}, \epsilon_{ijt})$ is *increasing* in its first argument. Each agent then chooses a certain product in a given time period, represented by $y_{ijt} = 1$, if and only if this product gives her the highest utility among all available products.

The infinite dimensionality of the terms u, A_{ij}, and ϵ_{ijt}, together with the model's additively non-separable interaction structure, jointly generates a rich class of unobserved heterogeneity. Across each agent-product combination ij, we are effectively allowing for nonparametric variations in agent utilities. Such variation proxies for the effects of complicated unobserved factors that influence choice behavior, such as brand loyalty, subtle flavors, and unique styles of products. In addition, we work with a nonparametric time homogeneity

assumption on the error terms ϵ_{ijt} that restricts ϵ_{ijt} and ϵ_{ijs} from two periods t and s to have the same marginal distribution given the observed covariates from the two periods. Apart from this, we impose no parametric restrictions on the distribution of ϵ_{ijt} and no additional restrictions on its dependence across time t and products j. In particular, the fully unrestricted dependence of ϵ_{ijt} across products j allows our framework to remain robust to the well-known "Blue-Bus/Red-Bus" problem and related pathologies that arise in many standard multinomial choice models, as discussed in Berry and Pakes (2007).[1]

The generality of our framework nests many semiparametric (and parametric) panel multinomial choice models with scalar fixed effects, scalar error terms, and varying degrees of additive separability, including the following standard specification:

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

Relative to existing work, our framework accommodates both the infinite dimensionality of unobserved heterogeneity and non–additive separability in agent utilities, under a standard time-homogeneity assumption on the idiosyncratic error term that is widely used in the literature.

Our identification strategy leverages multivariate monotonicity in contrapositive form. The intuition is straightforward: if the choice probability of a given product (or subset of products) strictly *increases* from one period to the next, then it *cannot* be that this product (or all products in the subset) becomes *worse* while all other products become *better* over the two periods. By applying this contraposition to a carefully constructed inequality in conditional choice probabilities, we obtain an identifying restriction on the index values that is free of all infinite-dimensional nuisance parameters. We further show that, in a two-period setting, the identified set obtained by aggregating these restrictions across all product subsets is sharp.

Based on our identification result, we provide consistent two-step set (or point) estimators, together with a computational algorithm adapted to the technical challenges of our

[1] See the discussion after Assumption 3 in Section 2.1 for more details.

framework. The first stage takes the form of a standard nonparametric regression, where we estimate a collection of intertemporal differences in conditional choice probabilities. In the second stage, we numerically minimize our sample criterion function with the first-stage estimates plugged in. A highlight of our computational procedure is the adoption of a spherical-coordinate reparameterization of our criterion functions in terms of *angles*, which enables us to exploit a combination of topological, geometric and computational advantages. A simulation study is conducted to analyze the finite-sample performance of our method and the adequacy of our computational procedure for practical implementation.

We also provide an empirical illustration of our procedure, where we use the NielsenIQ data on popcorn sales in the United States to explore the effects of marketing promotion. 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. Intuitively, marketing managers are more likely to promote products for which they know consumers are more sensitive to price and promotion. Hence, the average effective price sensitivities of promoted products tend to be larger than those not promoted due to the selection effect. Given the non-additive nature of such selection effects, estimators based on additive separability will be biased. In contrast, our method is robust to such confounding effects, thus producing more sensible estimates.

The validity of our identification strategy, as well as our estimation procedure, relies solely on monotonicity in an index structure, and thus extends naturally beyond panel multinomial choice models. We also introduce the *multi-index single-crossing (MISC)* condition framework, a general econometric framework under which our method can be applied. This framework encompasses the key ingredients of a large class of models, such as binary choice models with awareness, binary choice with endogeneity, dyadic network formation, bilateral matching, and endogenous censoring.

We acknowledge several limitations of our identification approach. First, our current

model setup does not allow for time-varying endogeneity between observed covariates X_{ijt} and the error term ϵ_{ijt}, which effectively rules out the inclusion of contemporaneously endogenous covariates and/or lagged outcomes. See a subsequent paper by Gao and Wang (2025) for a weaker version of the time homogeneity assumption that can be exploited for identification in panel multinomial choice models with endogenous and dynamic covariates. Second, our current approach does not allow for random coefficients on time-varying covariates. That said, rich forms of time-invariant taste heterogeneity have been absorbed into the nonparametric fixed effects under our current setting. Third, in the short-panel setting, our current approach effectively "differences out" the unobserved individual fixed effects A_i. As a result, it cannot identify counterfactual parameters that depend on the distribution of A_i. However, in long panels, our approach can be adapted to identify counterfactual parameters, which we discuss in more detail in Appendix E.

This paper builds upon and contributes to a large literature on semiparametric (and parametric) discrete choice models, dating back to McFadden (1974) and Manski (1975), and more specifically to the line of literature on panel multinomial choice models. Our work is most closely related to Pakes and Porter (2024), who also exploit weak monotonicity and time homogeneity, but restrict the effect of unobserved heterogeneity to be a scalar index that is additively separable from the index of observable characteristics. Shi, Shum, and Song (2018) exploit cyclical monotonicity of *vector*-valued functions in a fully additive panel multinomial choice model. Khan, Ouyang, and Tamer (2021) consider another additive model, but utilize the subsample of observations with time-invariant covariates along *all products but one* so as to leverage univariate monotonicity. Honoré and Kyriazidou (2000) also exploit univariate monotonicity when certain covariates across two periods are equal in a dynamic panel setting. Chernozhukov, Fernández-Val, and Newey (2019) consider a model with an additive effect under an "on-the-diagonal" restriction (i.e., when covariates at two different time periods coincide). By allowing non-additiveness in the specification of utility functions and infinite-dimensional fixed effects, our method is different from and thus

complementary to those proposed in these aforementioned papers.

This paper is also connected to the literature on the nonparametric identification and estimation in discrete choice models (e.g., Berry and Haile (2014); Compiani (2022)). That literature assumes monotonicity restrictions of the demand functions in product-market specific parametric indices to invert the demand system. This is particularly useful as it leads to a system of equations with only one unobservable per equation, from which the unobservable product-market specific demand shifter can be successfully constructed. Our paper considers a different model, but also leverages monotonicity restrictions in parametric indices to facilitate identification and estimation of the structural parameters. In both cases, the index assumption restricts the amount of unobserved heterogeneity in preferences for the variables included in the index.

More broadly, our work connects to the semiparametric literature on the identification and estimation of models characterized by monotonicity in a single parametric index. A related class of estimators that leverage univariate monotonicity, known as *maximum score* or *rank-order estimator*s, dates back to a series of important contributions by Manski (1975, 1985, 1987), and is further investigated in Han (1987), Horowitz (1992), Abrevaya (2000), Honoré and Lewbel (2002), Fox (2007), and Yan and Yoo (2019).[2] Despite the similar reliance on monotonicity, the *multi-index* nature of our model, and more importantly the multivariate monotonicity condition that we leverage, induce key differences from the *single-index* (and univariate monotonicity) setting, leading to a substantially different estimation method relative to rank-order estimators.

The rest of this paper is organized as follows. Section 2 introduces our main model specifications and assumptions. Section 2.2 presents our key identification strategy. In Section 3, we provide consistent estimators along with a computational procedure to implement it. Section 4 discusses the generalization of our method to the framework of multi-index

[2]We clarify that Manski (1975), Fox (2007) and Yan and Yoo (2019) consider multinomial choice models with multiple parametric indices, but focus on settings where the "multi-index" problem can be reduced to leverage single-variate monotonicity (or a single-variate rank-order property).

single-crossing conditions. Sections 5 and 6 switch back to our main panel multinomial choice model, for which we provide a simulation study and an empirical illustration using the NielsenIQ data. We conclude in Section 7.

2 Panel Multinomial Choice Model

2.1 Model and Assumptions

In this section, we present a semiparametric panel multinomial choice model featuring infinite-dimensional unobserved heterogeneity and flexible forms of non-separability, which serves as the main framework for illustrating our identification and estimation method.

Specifically, we consider the following model, in which individual i chooses product j at time t if and only if i prefers product j to all other alternatives at time t:

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

where:

- $i \in \{1, \ldots, N\}$ denotes N *individuals*.
- $j \in \mathcal{J} := \{1, \ldots, J\}$ denotes the set of J choice alternatives, with *products* indexed by $1, \ldots, J$. Throughout this paper, we treat the number of products J as fixed.
- $t \in \{1, \ldots, T\}$ denotes the $T \geq 2$ time periods. In this paper, we consider a short-panel setting in which T is fixed.
- X_{ijt} is an $\mathbb{R}^D$-valued vector of observable characteristics specific to each agent–product–time tuple ijt. These may include, for example, buyer characteristics such as income, product characteristics such as price and promotion status, as well as interaction and higher-order terms of these variables.
- y_{ijt} is an observable binary variable, with $y_{ijt} = 1$ indicating that buyer i chooses product j at time t, and $y_{ijt} = 0$ indicating otherwise.

- $\beta_0 \in \mathbb{R}^D$ is the finite-dimensional parameter of interest.
- A_{ij} represents an ij-specific time-invariant unobserved heterogeneity term of arbitrary dimension, which we refer to as the ij-specific *fixed effect*.
- ϵ_{ijt} is an ijt-specific unobserved error term of arbitrary dimension, which captures time-idiosyncratic utility shocks to product j for agent i at time t.
- u is an unknown function, interpreted as a *utility function* that aggregates the parametric index $X_{ijt}'\beta_0$, the fixed effect A_{ij}, and the error term ϵ_{ijt} into a scalar representing agent i's utility from choosing product j at time t.

We first present our main modeling assumptions and then discuss these assumptions in conjunction with our model specification (1). To economize on notation, we refer to the collection of variables concatenated along product and time dimensions: $\mathbf{X}_{it} = (X_{ijt})_{j=1}^J$, $\mathbf{X}_i = (\mathbf{X}_{it})_{t=1}^T$, $\mathbf{A}_i = (A_{ij})_{j=1}^J$, $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^J$, and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{it})_{t=1}^T$. We also write $\delta_{ijt} = X_{ijt}'\beta_0$ to denote the parametric index.

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

Assumption 1 is a standard assumption.[3] Recall that the number of time periods T is held fixed, and we focus on a short panel setting with cross-sectional asymptotics.

Assumption 2 (Monotonicity in the Index)**.** *For every realization of* (A_{ij}, ϵ_{ijt})*, the mapping* $\tilde{\delta} \longmapsto u(\tilde{\delta}, A_{ij}, \epsilon_{ijt})$ *is weakly increasing in the scalar-valued argument* $\tilde{\delta}$.

Essentially, Assumption 2 states that $u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$, the utility of individual i from choosing product j at time t, is weakly increasing[4] in the index δ_{ijt}. Given the index structure,

[3] It is worth noting that we have not yet imposed any explicit restrictions on the structure of the spaces in which the arbitrary-dimensional random elements $\mathbf{A}_i$ and $\boldsymbol{\epsilon}_i$ are defined. However, implicit in our model specification and in Assumption 1 is the requirement that $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \epsilon_i)$ be well-defined as random elements—that is, measurable functions—on a sufficiently rich probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

[4] It should be clarified that increasingness is without loss of generality given monotonicity, since the index $\delta_{ijt} = X_{ijt}'\beta_0$ contains an unknown parameter with unrestricted signs.

monotonicity itself is a relatively mild assumption. In the standard panel multinomial choice model with scalar-valued A_{ij} and ϵ_{ij} along with additive u, i.e., $u(\delta_{ijt}, A_{ij}, \epsilon_{ijt}) = \delta_{ijt} + A_{ij} + \epsilon_{ijt}$, Assumption 2 is trivially satisfied (with strictness) by construction.

In a way, Assumption 2 endows the index δ_{ijt} and the parameter β_0 with economic interpretations. Under Assumption 2, δ_{ijt} may be considered as a quality measure of the match between agent i and product j based on their observable characteristics at time t, inducing an interpretation of 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*. Hence, without Assumption 2, it would be hard to interpret β_0.

Moreover, the monotonicity restriction is imposed on δ_{ijt}, but not directly on any specific observable characteristics in X_{ijt}: quadratic or higher-order polynomial terms, and other functions of observable characteristics can be included in X_{ijt} whenever appropriate.

Assumption 3 (Pairwise Time Homogeneity)**.** *The distributions of* $\boldsymbol{\epsilon}_{it}$ *and* $\boldsymbol{\epsilon}_{is}$ *conditional on* $(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i)$ *across any pair of periods* $t \neq s \in \{1, \ldots, T\}$ *satisfy*

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

Assumption 3, a multinomial extension of the group homogeneity assumption in Manski (1987), is also a standard assumption in the literature on panel multinomial choice models, such as in Chernozhukov, Fernández-Val, and Newey (2019), Shi, Shum, and Song (2018), and Pakes and Porter (2024).[5] As shown in the next subsection, Assumption 3 is the key condition underlying our identification strategy. Essentially, we rely on Assumption 3 to

[5]In particular, Pakes and Porter (2024) investigate the following panel multinomial choice model:

$$y_{ijt} = \mathbb{1}\left\{g_j\left(X_{ijt}, \beta_0\right) + f_j\left(A_{ij}, \epsilon_{ijt}\right) > \max_{k \neq j} g_k\left(X_{ikt}, \beta_0\right) + f_k\left(A_{ik}, \epsilon_{ikt}\right)\right\}, \tag{2}$$

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 $g_j(X_{ijt}, \beta_0)$ and the unobserved heterogeneity index $f_j(A_{ij}, \epsilon_{ijt})$ is still maintained. Moreover, although the dimensions of A_{ij} and ϵ_{ijt} are not restricted in Pakes and Porter (2024), their joint effect is effectively summarized by a single scalar index $f_j(A_{ij}, \epsilon_{ijt})$. 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.

link (a particular class of) intertemporal changes in conditional choice probabilities between two periods s and t to intertemporal changes in the parametric indices of all products, $(\delta_{ijt} - \delta_{ijs})_{j \in \mathcal{J}}$, thereby yielding identifying restrictions for the parameter β_0. Note that Assumption 3 concerns only the marginal distributions of $\boldsymbol{\epsilon}_{it}$ in different periods, and we make no explicit assumptions about the serial dependence between $\boldsymbol{\epsilon}_{is}$ and $\boldsymbol{\epsilon}_{it}$, nor do we impose any restrictions on the dependence structure of A_{ij} and ϵ_{ijt} across products $j \in \mathcal{J}$.

It is worth emphasizing that the absence of restrictions on the cross-product dependence structure of ϵ_{ijt} in our setup also enables our choice model to circumvent the well-known "Blue-Bus/Red-Bus" problem[6] that arises in discrete choice models with additive errors that are assumed to be independent across products. Specifically, consider a standard multinomial logit model: if we arbitrarily create a new dummy product by replicating an existing one and giving it a different name, then under various (mixed) logit models a new independent copy of logit error is drawn for this new dummy product, which results in a strict increase in the consumers' indirect utilities, even though the new product is a simple duplicate of an existing product. This "Blue-Bus/Red-Bus" phenomenon also implies that consumers' indirect utilities would diverge to infinity if we keep adding such dummy new products, and that consumers will choose one of these duplicate products with increasingly higher probabilities, both of which are unrealistic. This problem, as well as other conceptual issues with independent additive errors, has been well discussed, say, in Berry and Pakes (2007). We emphasize that our current model does *not* lead to the "Blue-Bus/Red-Bus" problem, since we allow errors to be arbitrarily correlated across products. Hence, when duplicate products are created, the errors of duplicate products are allowed to be perfectly correlated (as they should be), so that adding duplicate products with different name labels does not blow up the indirect utility of the consumer, nor does it make the consumer more likely to choose one of the duplicates relative to any set of remaining products.

That said, Assumption 3 does entail certain limitations for the model. First, the assump-

[6]See, for example, McFadden (1974) and Train (2009) for descriptions of the Independence of Irrelevant Alternatives property and the "Blue-Bus/Red-Bus" problem.

tion effectively rules out time-varying endogeneity[7]: for example, if $\mathbf{X}_{it} \neq \mathbf{X}_{is}$, the conditional distribution of $\boldsymbol{\epsilon}_{it}$ is nevertheless assumed to be the same as that of $\boldsymbol{\epsilon}_{is}$, which is likely to be violated if the distribution of $\boldsymbol{\epsilon}_{it}$ covaries with that of $\mathbf{X}_{it}$. This is admittedly a limitation of the approach, though it is common to this line of literature that utilizes Assumption 3 as cited before. Nonetheless, subsequent work by Gao and Wang (2025) has proposed a weakened notion of the stationarity/homogeneity assumption that is only imposed on "exogenous covariates", and has proposed an approach for partial identification, albeit in a slightly different setting with additive scalar-valued fixed effects. Relatedly, Li (2024) proposes a correlated random coefficient linear panel model in which regressors can be correlated with time-varying, individual-specific random coefficients, thereby accommodating time-varying endogeneity in the covariates. Second, a further limitation of Assumption 3 is that it rules out random coefficients, a modeling device that was popularized by Berry, Levinsohn, and Pakes (1995) due to its ability to generate rich substitution patterns among products with multi-dimensional observable characteristics. 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. Our infinite-dimensional fixed-effect approach is thus 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.

Beyond Assumption 3, another limitation of the paper is that we treat the distribution of unobserved heterogeneity (i.e., the fixed effects) as a nuisance parameter and focus on the identification of β_0. Many counterfactual parameters require knowledge of the distribution of the relevant unobserved heterogeneity terms, which is not identified given that the fixed effects are "differenced out" under our current approach in a short-panel setting. To this end, we discuss in Appendix E how to use the estimated β_0 in a long panel setting ($T \to \infty$) to perform counterfactual analysis. The idea is when a long panel is available, we

[7]Note that time-invariant endogeneity can be incorporated through the unrestricted dependence between ϵ_{ijt} and the fixed effect A_{ij}, which can be arbitrarily correlated with $\mathbf{X}_i$ (in time-invariant manners).

can consistently estimate for each individual certain parameters of interest by using the observations only from that individual, which effectively controls for the unobserved $\mathbf{A}_i$.

Furthermore, there are interesting economic questions that can be addressed based on the knowledge of β_0, and we discuss two of them here.[8] First, consider research questions that focus on the existence and direction of an effect, which can often be determined by the relative magnitudes of effects from covariates as captured by β_0. For instance, one may ask whether advertising by a competitor within the same product category exerts a positive or negative influence on the demand for a focal brand. Theoretical considerations offer plausible arguments for both substitution and category-expansion effects (Simon and Arndt, 1980; Narayanan, Desiraju, and Chintagunta, 2004), rendering the sign of the net impact an empirical matter of interest. Second, the proposed approach may also be applied to model specification testing (or as a robustness check) for models with more restrictive specifications on the fixed effects and errors. Specifically, by comparing the estimate of β_0 obtained from our model with that from a more standard model (say, multinomial logit with fixed effects), one can assess whether the parametric restrictions, particularly those pertaining to unobserved heterogeneity, are supported by the data. This aligns with the broader econometric literature on specification testing (e.g., Hausman (1978); Vuong (1989)), and can be particularly useful for evaluating the validity of assumptions regarding the distributional structure or functional form of heterogeneity.[9]

2.2 Key Identification Strategy

In this section, we present our main semiparametric identification result for model (1) under Assumptions 2 and 3, and detail our key identification strategy, which exploits multivariate monotonicity in the presence of additive non-separability and nonparametric fixed effects.

[8] We thank an anonymous referee for the suggestions here.

[9] Ota and Otsu (2025) develop a specification test of parametric binary choice models via the maximum score estimator. Given that our approach generalizes the maximum score idea to settings with multivariate monotonicity and infinite-dimensional fixed effects, extending the specification testing idea of Ota and Otsu (2025) to our multivariate monotone panel setting can be a promising avenue for future work.

To start, fix any subset of products $\tilde{\mathcal{J}} \subseteq \mathcal{J}$, a pair of time periods $t \neq s \in \{1, \ldots, T\}$ and a generic realization of observable covariates in the two periods t and s, i.e., $(\mathbf{X}_{it}, \mathbf{X}_{is}) = (\mathbf{x}_t, \mathbf{x}_s) \in \text{Supp}(\mathbf{X}_{it}, \mathbf{X}_{is})$. For simpler notation, we write $\mathbf{X}_{i,ts} = (\mathbf{X}_{it}, \mathbf{X}_{is})$, $\mathbf{x}_{ts} = (\mathbf{x}_t, \mathbf{x}_s)$, $\delta_{jt} = x_{jt}^{'}\beta_0$, where x_{jt} denotes the j-th column of $\mathbf{x}_t$.

For each individual i, consider the intertemporal change in the probability of that individual choosing any product $j \in \tilde{\mathcal{J}}$ across periods t and s, conditional on $(\mathbf{X}_{i,ts}, \mathbf{A}_i)$, the joint realization of the fixed effect and the observable covariates in both periods t and s. Formally, writing $y_{i\tilde{\mathcal{J}}t} = \sum_{j\in\tilde{\mathcal{J}}} y_{ijt}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[y_{i\tilde{\mathcal{J}}t} - y_{i\tilde{\mathcal{J}}s} \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right] \\
&= \int \mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{jt}, A_{ij}, \epsilon_{ijt}\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{kt},\ A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{it} \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right) \\
&\quad - \int \mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{js}, A_{ij}, \epsilon_{ijs}\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{ks},\ A_{ik}, \epsilon_{iks}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{is} \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right) \\
&= \int \begin{bmatrix} \mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{jt}, A_{ij}, \tilde{\epsilon}_j\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{kt}, A_{ik}, \tilde{\epsilon}_k\right)\right\} \\ -\mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{js}, A_{ij}, \tilde{\epsilon}_j\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{ks}, A_{ik}, \tilde{\epsilon}_k\right)\right\} \end{bmatrix} \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}} \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right), \quad (3)
\end{aligned}
$$

where the last equality follows from Assumption 3 (pairwise time homogeneity):

$$\boldsymbol{\epsilon}_{is} \sim \boldsymbol{\epsilon}_{it} \sim \tilde{\boldsymbol{\epsilon}} \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i$$

in which $\tilde{\boldsymbol{\epsilon}}$ is a name-holder random element with the shared marginal distribution of $\boldsymbol{\epsilon}_{it}$ and $\boldsymbol{\epsilon}_{is}$ given $(\mathbf{x}_{ts}, \mathbf{A}_i)$.

Since u is weakly increasing in its index argument, the choice indicator

$$\mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{jt}, A_{ij}, \tilde{\epsilon}_j\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{kt}, A_{ik}, \tilde{\epsilon}_k\right)\right\}$$

is weakly increasing in the vector $(\delta_{jt})_{j\in\tilde{\mathcal{J}}}$ and decreasing in the remaining vector $(\delta_{kt})_{k\notin\tilde{\mathcal{J}}}$ for every possible realization of $\tilde{\boldsymbol{\epsilon}}$ and $\mathbf{A}_i$. Hence, if

$$\delta_{jt} \leq \delta_{js},\ \forall j \in \tilde{\mathcal{J}} \quad \text{and} \quad \delta_{kt} \geq \delta_{ks}\ \forall k \notin \tilde{\mathcal{J}}, \quad (4)$$

then, for every possible realization of $\tilde{\boldsymbol{\epsilon}}$ and $\mathbf{A}_i$, we have

$$\begin{bmatrix} \mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{jt}, A_{ij}, \tilde{\epsilon}_j\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{kt}, A_{ik}, \tilde{\epsilon}_k\right)\right\} \\ -\mathbb{1}\left\{\max_{j\in\tilde{\mathcal{J}}} u\left(\delta_{js}, A_{ij}, \tilde{\epsilon}_j\right) > \max_{k\notin\tilde{\mathcal{J}}} u\left(\delta_{ks}, A_{ik}, \tilde{\epsilon}_k\right)\right\} \end{bmatrix} \leq 0, \quad (5)$$

and consequently

$$\mathbb{E}\left[y_{i\tilde{\mathcal{J}}t} - y_{i\tilde{\mathcal{J}}s}\middle|\, \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right] \leq 0 \quad \text{for every possible realization of } \mathbf{A}_i. \tag{6}$$

Now, consider the observable intertemporal change in conditional choice probabilities:

$$\gamma_{\tilde{\mathcal{J}},ts}\left(\mathbf{x}_{ts}\right) := \mathbb{E}\left[y_{i\tilde{\mathcal{J}}t} - y_{i\tilde{\mathcal{J}}s}\middle|\, \mathbf{X}_{i,ts} = \mathbf{x}_{ts}\right]. \tag{7}$$

Then, whenever (4) holds, by (6) we can deduce that

$$\gamma_{\tilde{\mathcal{J}},ts}\left(\mathbf{x}_{ts}\right) = \int \underbrace{\mathbb{E}\left[y_{i\tilde{\mathcal{J}}t} - y_{i\tilde{\mathcal{J}}s}\middle|\, \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i\right]}_{\leq 0} \mathrm{d}\mathbb{P}\left(\mathbf{A}_i \middle|\, \mathbf{X}_{i,ts} = \mathbf{x}_{ts}\right) \leq 0.$$

In summary, since (4) implies inequality (5) for every possible realization of $\tilde{\boldsymbol{\epsilon}}$ and $\mathbf{A}_i$, this inequality will be preserved after $\tilde{\boldsymbol{\epsilon}}$ and $\mathbf{A}_i$ are integrated out *cross-sectionally* with respect to the conditional distribution $\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}, \mathbf{A}_i \middle|\, \mathbf{X}_{i,ts} = \mathbf{x}_{ts}\right)$, regardless of how complicated this unknown conditional distribution may be.

The next proposition formalizes the identification strategy described above, which produces an identifying restriction on the parameter β_0.

Proposition 1 (Key Identifying Restrictions)**.** *Under model* (1) *and Assumptions 1–3,*

$$\gamma_{\tilde{\mathcal{J}},ts}\left(\mathbf{x}_{ts}\right) > 0 \Rightarrow \text{NOT } \left\{\left(x_{jt} - x_{js}\right)' \beta_0 \leq 0,\ \forall j \in \tilde{\mathcal{J}} \text{ and } \left(x_{kt} - x_{ks}\right)' \beta_0 \geq 0\ \forall k \notin \tilde{\mathcal{J}}\right\} \tag{8}$$

for any (ordered) pair of time periods (t, s) *with* $t \neq s \in \{1, \ldots, T\}$*, any subset of products* $\tilde{\mathcal{J}} \subseteq \mathcal{J}$*, and any realization of observables* $\mathbf{x}_{ts} \in$ *Supp*$(\mathbf{X}_{i,ts})$*.*

Proposition 1 establishes an identifying restriction on β_0 that is free of all unknown nonparametric heterogeneity terms—u, $\mathbf{A}$, and $\boldsymbol{\epsilon}$—and holds in the presence of additive nonseparability and nonparametric fixed effects. Proposition 1 is also intuitive: if the total market share of the products in $\tilde{\mathcal{J}}$ increases between two periods, then it cannot be the case that the indices of products in $\tilde{\mathcal{J}}$ have all (weakly) worsened while those of products not in $\tilde{\mathcal{J}}$ have all (weakly) improved.

Theorem 1 (Identified Set)**.** *Let* B_0 *be the set of all* $\beta \in \mathbb{R}^D$ *such that* (8) *holds with* β *in lieu of* β_0*, for almost all* $\mathbf{x}_{ts} \in$ *Supp*$(\mathbf{X}_{i,ts})$*, all* $t \neq s \in \{1, \ldots, T\}$*, and all* $\tilde{\mathcal{J}} \subseteq \mathcal{J}$*. Then, under model* (1) *and Assumptions 1–3,* $\beta_0 \in B_0$*.*

We refer to B_0 as the *identified set*. In Appendix C, we provide sufficient conditions for point identification of β_0 up to a scale normalization. The assumptions we impose are similar to those used for point identification in the maximum-score literature, such as Manski (1985), and in related work on panel multinomial choice models, such as Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2021).

Relative to the well-known maximum-score criterion function studied by Manski (1985, 1987) under univariate monotonicity, our criterion function is non-standard. This non-standardness arises from a key distinction between multivariate and univariate monotonicity. To see this more clearly, consider the special case of a *single-index* setting ($J = 1$)[10], in which case the following equivalence relationship holds given the *univariate* monotonicity in the index:

$$\{\gamma(\mathbf{x}_{ts}) > 0\} \;\Leftrightarrow\; \left\{(x_t - x_s)' \beta > 0\right\}, \tag{9}$$

Such an "if-and-only-if" relationship is a unique feature of the single-index setting that *cannot* be generalized to the multi-index setting with $J \geq 2$, as the right-hand side of (8),

$$\text{NOT } \left\{(x_{jt} - x_{js})' \beta_0 \leq 0,\ \forall j \in \tilde{\mathcal{J}} \text{ and } (x_{kt} - x_{ks})' \beta_0 \geq 0\ \forall k \notin \tilde{\mathcal{J}}\right\},$$

does not imply $\gamma_{\tilde{\mathcal{J}},ts}(\mathbf{x}_{ts}) \geq 0$ in the converse direction. This breaks the "if-and-only-if" relationship that the maximum-score criterion function in Manski (1985, 1987) is built upon. Thus, the maximum-score estimator does not generalize to multi-index settings. The lack of "if-and-only-if" relationship in the multi-index setting leads to a key difference in the criterion functions, and consequently a different estimation approach. Importantly, while the original maximum score criterion and estimator cannot be generalized to multi-index settings, our procedure can be applied under a general econometric framework characterized by *multi-index single-crossing* conditions, which we introduce in Section 4.

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

2.3 Two-Period Sharpness

We now establish the sharpness of the identified set B_0 in a two-period setting. This sharpness result can be interpreted as the *pairwise sharpness* of the identifying restrictions in (8): for fixed periods s and t such that $s < t$, the inequality restrictions in (8) for (s, t) and (t, s) exhaust all the identifying information available from the model (its specification and assumptions) and from the distribution of the observable data in periods s and t.

Theorem 2 (Pairwise Sharpness)**.** *Under model* (1) *and Assumptions 1–3, B_0 is sharp for $T = 2$.*

The proof of Theorem 2 exploits and generalizes a corresponding result in Pakes and Porter (2024). Specifically, Pakes and Porter (2024) consider a specification where the utility index $X'_{ijt}\beta_0$ and an unobserved heterogeneity index $\lambda(A_{ijt}, \epsilon_{ijt})$ are additively separable, and establish the sharpness of their identification result by showing the existence of a nonnegative solution to a system of linear equations. Here, we consider a more general setup without requiring additive separability and propose a correspondingly more general identification argument. Nevertheless, we show that the sharpness of our identification result under our more general setup can be reduced to the nonnegative solvability of the same system of linear equations in Pakes and Porter (2024). Hence, by the result in Pakes and Porter (2024), our identification result is sharp.

Admittedly, Theorem 2 only establishes sharpness in a two-period setting; however, it does not directly imply "all-period" sharpness for $T \geq 3$. While the existence of an observationally equivalent latent error distribution can be established for any realization $\mathbf{X}_{i,ts}$ and any pair of periods (t, s) as in the proof of Theorem 2, to establish the stronger "all-period" sharpness result, we would need to show in addition that there exists an all-period joint distribution of latent errors that matches all-period observed joint conditional choice probabilities and satisfies the pairwise time homogeneity assumption. This appears to be a technically cumbersome exercise, given that the pairwise time-homogeneity assumption

enters as an implicit aggregate restriction on the two-period error distributions (with all other periods aggregated out). We thus do not pursue "all-period" sharpness here, and only present the sharpness result above as in Pakes and Porter (2024), which also focuses on two-period (pairwise) sharpness.

3 Estimation and Computation

3.1 Formulation of Population Criterion Function

We now propose a population criterion function that encodes the identifying information in Proposition 1. We represent the right-hand side of (8) in Boolean algebra by

$$\lambda_{\tilde{\mathcal{J}}}\left(\mathbf{x}_{ts};\beta\right) := \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k\in\tilde{\mathcal{J}}\}}\left(x_{kt}-x_{ks}\right)'\beta \geq 0\right\}, \tag{10}$$

where $(-1)^{\mathbb{1}\{k\in\tilde{\mathcal{J}}\}}$ takes the value -1 for $k \in \tilde{\mathcal{J}}$ and 1 for $k \notin \tilde{\mathcal{J}}$. Therefore, Proposition 1 can be written algebraically as: $\gamma_{\tilde{\mathcal{J}},t,s}\left(\mathbf{x}_{ts}\right) > 0$ implies $\lambda_{\tilde{\mathcal{J}}}\left(\mathbf{x}_{ts};\beta_0\right) = 0$ for any $\mathbf{x}_{ts} \in \text{Supp}\left(\mathbf{X}_{i,ts}\right)$.

We now define the following criterion function by taking a cross-sectional expectation over the random realization of $\mathbf{X}_{i,ts}$ and aggregating over all subsets $\tilde{\mathcal{J}} \subseteq \mathcal{J}$:

$$Q_{t,s}\left(\beta\right) := \sum_{\tilde{\mathcal{J}}\subseteq\mathcal{J}} \mathbb{E}\left[\mathbb{1}\left\{\gamma_{\tilde{\mathcal{J}},t,s}\left(\mathbf{X}_{i,ts}\right) > 0\right\}\lambda_{\tilde{\mathcal{J}}}\left(\mathbf{X}_{i,ts};\beta\right)\right], \tag{11}$$

which is nonnegative and minimized to zero at β_0. Without normalization and further assumptions for point identification, there could be multiple values of β that minimize $Q_{t,s}$ to zero.

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

$$Q^G_{t,s}\left(\beta\right) := \sum_{\tilde{\mathcal{J}}\subseteq\mathcal{J}} \mathbb{E}\left[G\left(\gamma_{\tilde{\mathcal{J}},t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_{\tilde{\mathcal{J}}}\left(\mathbf{X}_{i,ts};\beta\right)\right], \tag{12}$$

which is also minimized to zero at β_0. The sign-preserving function G, if further set to be

monotone, continuous, or bounded, serves as a *smoothing* function that can improve the finite-sample performance of our estimators. We provide more discussions on function G in the next section, when we construct estimators based on the sample analog of the population criterion function defined here.

$Q_{t,s}^G$ above is defined for a fixed pair of periods (t, s), but in practice we may utilize the information across all pairs of periods by defining the aggregated criterion function:

$$Q^G(\beta) := \sum_{t \neq s}^{T} Q_{t,s}^G(\beta), \quad \text{for any } \beta \in \mathbb{R}^D. \tag{13}$$

For notational simplicity, we suppress G in $Q_{t,s}^G$ and Q^G in the rest of this paper.

3.2 Two-Step Semiparametric Estimation

We construct our estimator as a semiparametric two-step M-estimator based on (13). The first stage of our procedure is concerned with nonparametrically estimating the intertemporal differences in conditional choice probabilities of the following form:

$$\gamma_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts}) = \sum_{j \in \tilde{\mathcal{J}}} \gamma_{j,t,s}(\mathbf{x}_{ts}),$$

where $\gamma_{j,t,s}(\mathbf{x}_{ts}) = \mathbb{E}[y_{ijt} - y_{ijs} | \mathbf{X}_{i,ts} = \mathbf{x}_{ts}]$ can be separately estimated for each product $j \in \mathcal{J}$.[11] We note that the first stage estimation includes the observable characteristics of all products J. For example, when $J = 3$ and $D = 3$, there are $3 \times 3 \times 2 = 18$ variables in the conditioned set of γ. Given the potentially large number of regressors, one may want to use neural networks (Bach, 2017; Chen and White, 1999) or penalized sieves (Chen, 2013) for the first-step estimation.

Given the first-stage estimators $\hat{\gamma}_{j,t,s}$ and the smoothing function G, in the second stage

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

we numerically compute minimizers of the sample criterion function,

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

It is worth noting that while $\hat{Q}_{t,s}(\beta)$ is defined as a summation over all 2^J possible subsets $\tilde{\mathcal{J}} \subseteq \mathcal{J}$, computationally there is no need to fully evaluate $Q_{\tilde{\mathcal{J}},t,s}(\beta)$ for each possible $\tilde{\mathcal{J}} \subseteq \mathcal{J}$ under a given β when the knife-edge cases of $(X_{ijt} - X_{ijs})'\beta = 0$ are ignorable.[12] This is because, as long as $\lambda_{\tilde{\mathcal{J}}}(\mathbf{X}_{i,ts}; \beta) = 0$, the contribution from the $\tilde{\mathcal{J}}$-summand would be zero. However, a careful inspection of $\lambda_{\tilde{\mathcal{J}}}$ reveals that $\lambda_{\tilde{\mathcal{J}}}(\mathbf{X}_{i,ts}; \beta) = 1$ only if $\tilde{\mathcal{J}} = \left\{j \in \mathcal{J} : (X_{ijt} - X_{ijs})' \beta \leq 0\right\}$. Hence, in practical implementation we may simply compute

$$\hat{Q}_{t,s}(\beta) := \frac{1}{N} \sum_{i=1}^{N} G\left(\sum_{j \in \mathcal{J}} \hat{\gamma}_{j,t,s}(\mathbf{X}_{i,ts}) \mathbb{1}\left\{(X_{ijt} - X_{ijs})' \beta \leq 0\right\}\right).$$

In addition, the scale of β_0 is not identified since $\lambda_j(\mathbf{X}_{i,ts}; \beta)$ consists of indicator functions of the form $\mathbb{1}\left\{(X_{ijt} - X_{ijs})' \beta \geq 0\right\}$. Hence, we impose the scale normalization $\beta_0 \in \mathbb{S}^{D-1} := \left\{v \in \mathbb{R}^D : \|v\| = 1\right\}$. Following Chernozhukov, Hong, and Tamer (2007), we define the set estimator by

$$\hat{B}_{\hat{c}} := \left\{\beta \in \mathbb{S}^{D-1} : \ \hat{Q}(\beta) \leq \min_{\tilde{\beta} \in \mathbb{S}^{D-1}} \hat{Q}\left(\tilde{\beta}\right) + \hat{c}\right\} \tag{14}$$

with $\hat{c} := O_p(c_N \log N)$.

We now introduce assumptions for establishing the consistency of $\hat{B}_{\hat{c}}$.

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

(i) $\gamma_{j,t,s} \in \Gamma$*, and* $\mathbb{P}(\hat{\gamma}_{j,t,s} \in \Gamma) \to 1$*, with* Γ *being a* $\mathbb{P}$*-Donsker class of functions in* $L_2(\mathbf{X})$*.*

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

Through Assumption 4 we take as given the large set of theoretical results on nonparametric regression in the literature. Many kernel-based and sieve-based methods have been devel-

[12] There are at least two reasons why the "knife-edge" events of the form $(X_{ijt} - X_{ijs})'\beta = 0$ should be ignored. First, $(X_{ijt} - X_{ijs})'\beta = 0$ technically occurs with probability zero for all β provided that $X_{ijt} \neq X_{ijs}$ almost surely, which is a natural assumption given that we require X_{ijt} be time-varying. Second, for programming reasons, it is often practically necessary to ignore knife-edge strict equalities of continuously valued variables, since such equalities are extremely sensitive to unavoidable numerical errors induced by the "machine epsilon."

oped, with their properties demonstrated under various sets of conditions. See Wasserman (2006) and Chen (2007) for more comprehensive surveys.

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

Assumption 5 is stronger than necessary for consistency per se given that our identification result is valid with any choice of the one-sided sign-preserving function G, nevertheless we take G to be Lipschitz to simplify the proof.

To state the next assumption, we decompose each row (corresponding to each product) of $\mathbf{x}_t - \mathbf{x}_s$ as the product of its norm and its *direction*, i.e., $\mathbf{x}_{jt} - \mathbf{x}_{js} \equiv r_j(\mathbf{x}_t - \mathbf{x}_s)\, v_j(\mathbf{x}_t - \mathbf{x}_s)$, where $r_j(\mathbf{x}_t - \mathbf{x}_s) := \|\mathbf{x}_{jt} - \mathbf{x}_{js}\|$, and $v_j(\mathbf{x}_{jt} - \mathbf{x}_{js}) := (\mathbf{x}_{jt} - \mathbf{x}_{js}) / \|\mathbf{x}_{jt} - \mathbf{x}_{js}\|$ if $\mathbf{x}_{jt} \neq \mathbf{x}_{js}$ while $v_j(\mathbf{x}_{jt} - \mathbf{x}_{js}) := \mathbf{0}$ if $\mathbf{x}_{jt} = \mathbf{x}_{js}$.

Assumption 6 (Continuous Distribution of Directions)**.** *The marginal distribution of* $v_j(\mathbf{X}_{it} - \mathbf{X}_{is})$ *has no mass point except possibly at* $\mathbf{0}$ *and is not supported on any proper linear subspace of* $\mathbb{R}^D$ *for each* (j, t, s) *tuple.*

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

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

Theorem 3 (Consistency)**.** *Under Assumptions 1–6, the set estimator* $\hat{B}_{\hat{c}}$ *is consistent in Hausdorff distance:* $d_H\left(\hat{B}_{\hat{c}}, B_0\right) = o_p(1)$, *where* $d_H\left(\hat{B}_{\hat{c}}, B_0\right) = \max\left\{\sup_{\beta \in \hat{B}_{\hat{c}}} \inf_{\tilde{\beta} \in B_0} \left\|\beta - \tilde{\beta}\right\|, \sup_{\beta \in B_0} \inf_{\tilde{\beta} \in \hat{B}_{\hat{c}}} \left\|\beta - \tilde{\beta}\right\|\right\}$. *Furthermore, if* β_0 *is point-identified on* $\mathbb{S}^{D-1}$, $\left\|\hat{\beta} - \beta_0\right\| = o_p(1)$ *for any* $\hat{\beta} \in \hat{B}_{\widehat{c}=0}$.

3.3 Computation

We now explain how we implement the semiparametric two-step estimation procedure proposed above. Since the model's criterion function is possibly non-convex, standard gradient-based optimizers are susceptible to converging to local minima. To address this, we employ a multi-stage adaptive-grid search algorithm that explores the parameter space more robustly and aims to locate the global minimizer of the objective function. The code for our computation algorithm is publicly available on GitHub.[13]

Choice of the Smoothing Function G

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

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

On the other hand, when $\gamma_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts})$ is positive and large so that $\mathbb{1}\left\{\gamma_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts}) > 0\right\}$ can be estimated well, the magnitude of $\gamma_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts})$ itself does not provide additional identifying information. By setting G to be bounded from above, we dampen the influence of large values of $\gamma_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts})$, so that the numerical minimization of $\hat{Q}$ is less sensitive to potentially large but redundant variations in $\hat{\gamma}_{\tilde{\mathcal{J}},t,s}(\mathbf{x}_{ts})$.

[13] https://github.com/mingliecon/GL_PMC

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

To minimize $\hat{Q}(\beta)$ over $\beta \in \mathbb{S}^{D-1}$, we work with a reparameterization of $\mathbb{S}^{D-1}$ with $D-1$ angles in spherical coordinates.[14] Specifically, define the angle space Θ by

$$\Theta := [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}, \tag{15}$$

and the transformation $\theta \longmapsto \beta(\theta)$ by standard spherical coordinate transformation. We now instead solve the optimization of $\hat{Q}(\beta(\theta))$ over Θ, which we further equip with its natural geodesic metric $\rho_\Theta\left(\theta, \tilde{\theta}\right) := \arccos\left(\beta(\theta)'\beta\left(\tilde{\theta}\right)\right)$. Note that $\rho_\Theta\left(\theta, \tilde{\theta}\right)$ is strongly equivalent[15] to the (imported) Euclidean distance $\left\|\beta(\theta) - \beta\left(\tilde{\theta}\right)\right\|$.

This reparameterization (Θ, ρ_Θ) enables us to exploit the compactness and convexity of the parameter space $\Theta = [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}$, which takes the form of a hyper-rectangle. First, (Θ, ρ_Θ) preserves all topological structures of the unit sphere, and particularly inherits the compactness of $\left(\mathbb{S}^{D-1}, \|\cdot\|\right)$, automatically satisfying the compactness condition usually imposed for extremum estimation and making it numerically feasible to initiate a grid on the whole parameter space. Second, while the unit sphere $\mathbb{S}^{D-1}$ is not convex, the new parameter space Θ becomes convex algebraically, making it computationally easy to define bisection points in the parameter space. Third, (Θ, ρ_Θ) 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, $\rho_\Theta\left((\pi - \epsilon, \theta_2, \ldots, \theta_{D-1}), (-\pi, \theta_2, \ldots, \theta_{D-1})\right) \to 0$ as $\epsilon \to 0$. This seemingly trivial property is nevertheless important in defining and interpreting whether certain parameter estimates converge asymptotically or not.

[14] The idea and the motivation for using the angle-space reparameterization can also be found in Manski and Thompson (1986), who however use only one angle parameter.

[15] Two metrics d_1 and d_2 defined on some nonempty set X are *strongly equivalent* if and only if there exist positive constants c_1 and c_2 such that $c_1 d_1(x, y) \leq d_2(x, y) \leq c_2 d_1(x, y)$ for every $x, y \in X$.

An Adaptive-Grid Algorithm

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

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

$$\underline{\Theta}^{(1)} := \left\{\theta \in \Theta^{(1)} : \ \hat{Q}(\theta) \leq \text{quantile}_\alpha\left(\hat{Q}\left(\Theta^{(1)}\right)\right)\right\}. \tag{16}$$

Step 3: Take the enclosing rectangle of $\underline{\Theta}^{(1)}$, by defining $\underline{\theta}_d^{(1)} := \min^* \underline{\Theta}_d^{(1)}$ and $\overline{\theta}_d^{(1)} := \max^* \underline{\Theta}_d^{(1)}$, where $\underline{\Theta}_d^{(1)} := \left\{\theta_d : \theta \in \underline{\Theta}^{(1)}\right\}$ for each $d = 1, \ldots, D-1$ and the operator $\min^*$ and $\max^*$ have standard definitions of min and max except for the first dimension $d = 1$. For the first dimension, it is necessary to account for the underlying spherical geometry and the periodicity of angles, i.e. $\theta_1 + 2\pi \equiv \theta_1$ and in particular $-\pi \equiv \pi$. This, however, is largely a programming nuisance: whenever $\underline{\Theta}_1^{(1)} \subsetneq \Theta_1^{(1)}$ crosses over at $-\pi$ and 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 1.

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

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

Note that the above is simply a sketch of our algorithm: see Appendix D and the documentation on GitHub for more implementation details.[16] To be conservative, we add in

[16]Our algorithm relies heavily on the compactness and convexity of the angle space Θ. Compactness

buffers at each step of refinement, keep track of both outer and inner boundaries of the lower-quantile set $\underline{\Theta}^{(m)}$, and make sure that the minimizers of the criterion functions at all computed points are indeed enclosed by the set returned in the end. We find the current algorithm to be conservative and to perform well in our simulations.

This multi-stage approach is designed to balance computational feasibility with a robust search. The initial coarse search efficiently discards large, suboptimal regions of the parameter space, while the subsequent refinement and boundary identification stages provide a high-precision estimate in the most promising area. However, the algorithm's performance is inherently tied to the selection of tuning parameters, particularly the initial grid size `M_Step` and the quantile used for pruning, which must be chosen carefully to ensure the global minimum is not discarded prematurely. Furthermore, it can get computationally intensive when the dimension of β is high. We find that running 1,000 simulations for $D = 3$ usually takes a few hours on modern computers, while for $D = 4$ it may take up to a day.

4 General Econometric Framework of Multi-Index Single-Crossing Conditions

Our key identification strategy, and consequently the associated estimation method, apply more widely beyond panel multinomial choice models. We now introduce a general econometric framework defined by *multi-index single-crossing* (MISC) conditions, and show how our proposed methods can be exploited in a wide range of models nested under the MISC condition framework.

Formally, let $(y_i, X_i)_{i=1}^n$ be a random sample of data with X_i distributed on the support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and y_i distributed on $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Let $h_0 : \mathcal{X} \to \mathbb{R}$ denote a functional of the conditional distribution of y_i given X_i that is directly identified from data. For each of

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 enables us to conveniently refine the grid by separately cutting each coordinate of $\overline{\Theta}^{(m)}$ into smaller pieces through simple division.

$j = 1, \ldots, J \in \mathbb{N}$, let $\phi_j : \mathcal{X} \to \mathbb{R}^{d_{\theta_j}}$ be some known transformation of X_i, and define $W_{ij} := \phi_j(X_i)$ with $W_i := (W_{i1}, \ldots, W_{iJ})$. Let $\theta_{0j} \in \Theta_j \subseteq \mathbb{R}^{d_{\theta_j}}$ be an unknown finite-dimensional parameter and write $\theta_0 := \left(\theta_{01}', \ldots, \theta_{0J}'\right)' \in \Theta := \times_{j=1}^{J} \Theta_j$.

Definition 1 (*Multi-Index Single-Crossing Condition*)**.** We say that (h_0, θ_0) satisfy the (*weak*) *multi-index single-crossing condition* if, for any realization $x \in \mathcal{X}$ and $w = \phi(x)$,

$$\begin{aligned} w_j'\theta_{0j} \geq 0,\ \forall j = 1, \ldots, J \quad &\Rightarrow \quad h_0(x) \geq 0, \\ w_j'\theta_{0j} \leq 0,\ \forall j = 1, \ldots, J \quad &\Rightarrow \quad h_0(x) \leq 0. \end{aligned} \tag{17}$$

The condition is said to be strict if the inequalities on the right-hand side of (17) are strict.

In words, the MISC condition states that if all the J parametric indices $w_1'\theta_{01}$, $w_2'\theta_{02}$, ..., and $w_J'\theta_{0J}$ are (weakly) positive, then the functional h_0 must be (weakly) positive; if the J indices are all zero, then h_0 must be zero; if the J indices are all negative, then h_0 must be negative. Essentially, the MISC condition provides a parsimonious way to semiparametrically model how the multiple economic factors jointly affect a certain statistic of the relevant economic outcome. The MISC condition basically requires that, if all the relevant factors reach certain thresholds, then the outcome statistics must also reach certain thresholds. Such requirements are often easy to obtain in an economic or econometric model: while multiple factors in an economic model may interact with each other in potentially complicated manners and there might be many configurations of the factors that lead to ambiguous theoretical predictions, there are also often simple configurations that we understand reasonably well. Hence, the MISC condition imposes only mild requirements on the underlying economic or econometric model for the problem and thus provides a general framework for semiparametric econometric analysis, in which most of the modeling ingredients can be left nonparametric except for the parametric indices that capture different economic factors in the problem.

Clearly, the panel multinomial choice model considered in previous sections falls under the MISC condition framework. Specifically, focusing on a pair of time periods (t, s) and a particular product j_0 for illustration, define $\theta_{0j} := \beta_0$, $h_0(\mathbf{X}_i) := \gamma_{j_0,ts}(\mathbf{X}_i)$, $W_{ij_0} :=$

$X_{ij_0t} - X_{ij_0s}$ and $W_{ij} := -(X_{ijt} - X_{ijs})$ for $j \neq j_0$. Then, the MISC condition (17) is satisfied under model (1) and Assumptions 1–3.

We now provide a few more examples of models nested in the MISC condition framework.

Example 1 (Binary Choice with Awareness)**.** Consider the following binary choice model

$$y_i = \mathbb{1}\left\{X_{i1}'\theta_{01} \geq u_i\right\} \cdot \mathbb{1}\left\{X_{i2}'\theta_{02} \geq v_i\right\}$$

where y_i denotes whether consumer i purchases a certain product or not, X_{i1} denotes a vector of covariates that influences the consumer's utility from a product, and X_{i2} denotes a vector of covariates that affects the consumer's awareness of the product (e.g., advertising). Here, we have $J = 2$, $X_i := (X_{i1}, X_{i2})$, $W_{i1} := X_{i1}$, and $W_{i2} := X_{i2}$. Define $h_0(x) := \mathbb{E}[y_i | X_i = x] - \frac{1}{4}$. Then, under the conditional median restrictions $\text{med}(u_i | X_i) = \text{med}(v_i | X_i) = 0$ and the conditional independence restriction $u_i \perp v_i | X_i$, it is true that

$$\begin{aligned} X_{i1}'\theta_{01} > 0,\ X_{i2}'\theta_{02} > 0 \quad &\Rightarrow \quad h_0(X_i) > 0, \\ X_{i1}'\theta_{01} < 0,\ X_{i2}'\theta_{02} < 0 \quad &\Rightarrow \quad h_0(X_i) < 0, \end{aligned}$$

satisfying the MISC condition.

Example 2 (Binary Choice with Endogeneity)**.** Consider the binary choice model

$$Y_i = \mathbb{1}\left\{W_i'\beta_0 \geq \epsilon_i\right\},$$

and let one component of W_i, say, W_{i1} be endogenous. Suppose that there exists a vector of instrumental variables Z_i and define $\xi_i := W_{i1} - Z_i'\gamma_0$ as the residual from the reduced-form linear projection of W_{i1} on Z_i. Assume that the endogeneity between ϵ_i and W_{i1} is captured by the following control function

$$\text{med}(\epsilon_i | Z_i, \xi_i) = \lambda(\alpha_0 \xi_i),$$

where λ is an unknown increasing function with location normalization $\lambda(0) = 0$, and the sign parameter $\alpha_0 \in \{-1, 1\}$ controls the direction of the monotonicity. The above can be viewed as an adaptation of the binary choice model that combines the conditional median restriction in Manski (1975) with the control function approach in Blundell and Powell (2004): here

we only impose the control function restriction on the conditional median instead of the whole distribution as in Blundell and Powell (2004). Then, writing $\overline{Z}_i := (W_{i1}, Z_i)$ and $\overline{\gamma}_0 := (-\alpha_0, \alpha_0\gamma_0)'$, we have

$$W_i'\beta_0 > 0,\ \overline{Z}_i'\overline{\gamma}_0 > 0 \quad \Rightarrow \quad \mathbb{E}\left[Y_i \middle| W_i, Z_i\right] > \frac{1}{2}$$

and its "<" counterpart, which can be viewed as a MISC condition with $K = 2$, $h_0(W_i, Z_i) := \mathbb{E}\left[Y_i - \frac{1}{2}\middle| W_i, Z_i\right]$, $\phi_1(W_i, Z_i) := W_i$, $\phi_2(W_i, Z_i) := (W_{i1}, Z_i)$, $\theta_{0,1} := \beta_0$, and $\theta_{0,2} := \overline{\gamma}_0$.

Example 3 (Dyadic Network Formation)**.** Consider the dyadic network formation model of Gao, Li, and Xu (2023), which extends Graham (2017) to a semiparametric setting:

$$\mathbb{E}\left[y_{ij} \middle| X_i, X_j, A_i, A_j\right] = \psi\left(w(X_i, X_j)'\theta_0, A_i, A_j\right).$$

Here y_{ij} is a binary outcome indicating whether individuals i and j are linked in an undirected network, X_i and X_j are the individuals' observable covariates, $w(X_i, X_j)$ is a known pairwise transformation of individual covariates (with the leading example being $w_h(X_i, X_j) := |X_{i,h} - X_{j,h}|$ for each coordinate $h = 1, \ldots, d_x$), A_i and A_j are unobserved individual degree heterogeneity terms, and $\psi : \mathbb{R}^3 \to \mathbb{R}$ is an unknown function assumed to be increasing in all its three arguments. Specifically, fixing a particular pair of individuals $(\bar{i}, \bar{j})$ and two realizations $\overline{x}, \underline{x}$ of X_i, it can be shown that, with

$$\overline{w} := w\left(x_{\bar{j}}, \overline{x}\right) - w\left(x_{\bar{i}}, \overline{x}\right), \quad \underline{w} := w\left(x_{\bar{i}}, \underline{x}\right) - w\left(x_{\bar{j}}, \underline{x}\right),$$

and

$$\begin{aligned} h_0(\overline{x}, \underline{x}) := &\max\left(0, \mathbb{E}\left[y_{\bar{i}k} - y_{\bar{j}k}\middle| X_k = \overline{x}\right]\right)\mathbb{E}\left[y_{\bar{i}k} - y_{\bar{j}k}\middle| X_k = \underline{x}\right] \\ &- \max\left(0, \mathbb{E}\left[y_{\bar{j}k} - y_{\bar{i}k}\middle| X_k = \overline{x}\right]\right)\mathbb{E}\left[y_{\bar{j}k} - y_{\bar{i}k}\middle| X_k = \underline{x}\right] \end{aligned}$$

the weak MISC condition is satisfied under mild conditions:

$$\begin{aligned} \overline{w}'\theta_0 > 0,\ \underline{w}'\theta_0 > 0 \quad &\Rightarrow \quad h_0(\overline{x}, \underline{x}) \geq 0, \\ \overline{w}'\theta_0 < 0,\ \underline{w}'\theta_0 < 0 \quad &\Rightarrow \quad h_0(\overline{x}, \underline{x}) \leq 0. \end{aligned}$$

Example 4 (Censored Monotone Transformation Model with Endogeneity)**.** The approach proposed in Example 2 above can also be adapted to the following censored monotone trans-

formation model with endogeneity:

$$Y_i = \max\left\{\phi\left(W_i'\beta_0, \epsilon_i\right), 0\right\},$$

where one component of the observed covariates, W_{i1}, is endogenous, and ϕ is an unknown bivariate increasing function. This model generalizes the usual censored regression model $Y_i = \max\left\{W_i'\beta_0 + \epsilon_i, 0\right\}$, say, in Blundell and Powell (2007), by incorporating a flexible unknown monotone transformation ϕ with non-additive error term. Since β_0, ϵ_i and ϕ are all unknown, one may normalize $\phi(0,0) = 0$. By the equivariance of (conditional) quantiles under monotone transformations, we have

$$\operatorname{med}(Y_i | W_i, Z_i) = \max\left\{\phi\left(W_i'\beta_0, \operatorname{med}(\epsilon_i | W_i, Z_i)\right), 0\right\}.$$

Similar to Example 2, define $\xi_i := W_{i1} - Z_i'\gamma_0$ as the residual from the reduced-form linear projection of W_{i1} on the instrumental variables Z_i, and assume that the endogeneity between ϵ_i and W_{i1} is captured by the control function $\operatorname{med}(\epsilon_i | W_i, Z_i) = \operatorname{med}(\epsilon_i | Z_i, \xi_i) = \lambda(\alpha_0\xi_i)$, where λ is an increasing function with normalization $\lambda(0) = 0$. Writing $\overline{Z}_i := (W_{i1}, Z_i)$ and $\overline{\gamma}_0 := (\alpha_0, -\alpha_0\gamma_0)'$, we have

$$\begin{aligned} W_i'\beta_0 > 0,\ \overline{Z}_i'\overline{\gamma}_0 > 0 \quad &\Rightarrow \quad \operatorname{med}(Y_i | W_i, Z_i) > 0 \text{ and} \\ W_i'\beta_0 \leq 0,\ \overline{Z}_i'\overline{\gamma}_0 \leq 0 \quad &\Rightarrow \quad \operatorname{med}(Y_i | W_i, Z_i) = 0, \end{aligned}$$

which can be viewed as a MISC condition with a weak "$\leq$" side and $h_0(X_i) := \operatorname{med}(Y_i | W_i, Z_i)$ given by the conditional median function.

Based on the MISC conditions (17), we can again obtain identifying restrictions by taking their logical contrapositions, which can be encoded algebraically in a similar way as in (1) and (12). Specifically, let G be a one-sided sign-preserving function as in (12) and define

$$\lambda(W_i; \theta) := \prod_{j=1}^{J} \mathbb{1}\left\{W_{ij}'\theta_j \leq 0\right\}.$$

Proposition 2. *Under condition (17), we have*

$$\begin{aligned} h_0(X_i) > 0 \ &\Rightarrow \ NOT\ \left\{W_{ij}'\theta_j \leq 0\ \forall j\right\}, \\ h_0(X_i) < 0 \ &\Rightarrow \ NOT\ \left\{W_{ij}'\theta_j \geq 0\ \forall j\right\}. \end{aligned}$$

Furthermore, with $Q(\theta) := Q_+(\theta) + Q_-(\theta)$ *where*

$$Q_+(\theta) := \mathbb{E}\left[G\left(h_0(X_i)\right)\lambda(W_i;\theta)\right] \quad \textit{and} \quad Q_-(\theta) := \mathbb{E}\left[G\left(-h_0(X_i)\right)\lambda(-W_i;\theta)\right],$$

we have $Q(\theta) \geq Q(\theta_0) = 0$.

Proposition 2 generalizes Theorem 1. Notice that Proposition 2 applies to all functionals h_0 of the conditional distribution of y_i given $\mathbf{X}_i$ that satisfy the MISC conditions.

One could also proceed with the two-step estimation procedure described in Section 3. Given a first-stage nonparametric estimator $\hat{h}$ of h_0, we can estimate θ_0 (or the identified set) by minimizing the sample criterion $\hat{Q}(\theta) := \hat{Q}_+(\theta) + \hat{Q}_-(\theta)$ with

$$\hat{Q}_+(\theta) := \frac{1}{n}\sum_{i=1}^{n} G\left(\hat{h}(X_i)\right)\lambda(W_i;\theta) \quad \text{and} \quad \widehat{Q}_-(\theta) := \frac{1}{n}\sum_{i=1}^{n} G\left(-\hat{h}(X_i)\right)\lambda(-W_i;\theta).$$

5 Simulation

We now switch back to the panel multinomial choice model introduced in Section 2 and examine the finite sample performance of our proposed estimator. For each DGP, we run $M = 1,000$ simulations of model (1) with the following utility specification:

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

in which A_{i0} is an unobserved scale fixed effect that captures agent-level heteroskedasticity in utilities, and A_{ij} is an unobserved location shifter specific to each agent-product pair. The ability to deal with nonlinear dependence caused by unobserved fixed effects in a relatively robust way is a distinctive feature of our method compared with existing approaches. To allow for such dependence, we generate correlation between the observable characteristics $\mathbf{X}_i$ and the fixed effects $\mathbf{A}_i$ via a latent variable Z. We draw $Z_i \sim \mathcal{N}(0,1)$ and let $A_{i2} = [Z_i]_+$. We construct $X_{ijt,2} = W_{ijt} + Z_i$ with $W_{ijt} \sim \mathcal{N}(0, 2J)$. Thus, X and A are correlated via Z. The DGPs for the rest of $\mathbf{A}$ and $\mathbf{X}$ are: $A_{i0} \sim \mathcal{U}[2, 2.5]$, $A_{i1} \equiv 0$, $A_{ij} \sim \mathcal{U}[-0.25, 0.25]$ for $j \geq 3$, $X_{ijt,1} \sim \mathcal{U}[-1, 1]$, $X_{ijt,d} \sim \mathcal{N}(0,1)$ for $d \geq 3$. Furthermore, we set $\overline{\beta}_0 = (2, 1, \ldots, 1)' \in \mathbb{R}^D$ and $\beta_0 = \overline{\beta}_0 / \left\|\overline{\beta}_0\right\|$, and draw $\epsilon_{ijt} \sim TIEV(0,1)$. To summarize, for each

of the $M = 1,000$ simulations we first generate $(\beta_0, \mathbf{X}_{it}, \mathbf{A}_i, \boldsymbol{\epsilon}_{it})$ for all (i,t) pairs. Then, we calculate the individual choice $\mathbf{Y}$ matrix according to model (1). Next, we compute $\hat{\beta}$ from the simulated observable data of $(\mathbf{X}, \mathbf{Y})$. To obtain $\hat{\beta}$, we first use nonparametric regression with second-order polynomial basis functions with ℓ_1-regularization and 10-fold cross validation to estimate γ. Then, we apply the adaptive-grid algorithm detailed in Section 3.3. Finally, we assess how well $\hat{\beta}$ performs compared with the true value β_0.[17]

Baseline Results

For the baseline configuration, we set $N = 10,000$, $D = 3$, $J = 3$, and $T = 2$. Since in this case the conditions for the point identification are satisfied, any point from the argmin set $\hat{B}_b := \arg\min_{\beta \in \mathbb{S}^{D-1}} \hat{Q}_b(\beta)$ is a consistent estimator of β_0 for each round of simulation $b = 1, \ldots, M$. Specifically, we define

$$\hat{\beta}^u_{b,d} := \max \hat{B}_{b,d}, \quad \hat{\beta}^l_{b,d} := \min \hat{B}_{b,d}, \quad \text{and} \quad \hat{\beta}^m_{b,d} := \frac{1}{2}\left(\hat{\beta}^u_{b,d} + \hat{\beta}^l_{b,d}\right),$$

where $\hat{\beta}^u_{b,d}$, $\hat{\beta}^l_{b,d}$, and $\hat{\beta}^m_{b,d}$ represent the maximum, minimum, and middle point along dimension d for each round of simulation b of the argmin set $\hat{B}$, respectively.

[Table 1 about here.]

Table 1 summarizes our baseline results. In the first row we use the middle point $\hat{\beta}^m$ along each dimension of $\hat{B}$ to calculate the bias. The biases are very small across all three dimensions with a magnitude between -0.0034 and -0.0005. The next two rows show the biases in estimating $\beta_{0,d}$ using $\hat{\beta}^u_d$ and $\hat{\beta}^l_d$ respectively, which are again close to zero. The fourth row reports the average widths of the set $\hat{B}$ along each dimension. These widths are small relative to the magnitude of β_0. The fifth and sixth rows summarize the standard

[17] In Appendix F, we provide additional simulation results. Specifically, we first present a graphical illustration of the identified set B_0 based on the population criterion (13). Second, we inspect how our estimator performs without point identification. Third, we vary (D, J, T) to examine how robust our method is against various simulation specifications. Lastly, we include a simulation illustration of the robustness of our approach to the "Blue-Bus/Red-Bus" problem.

deviation and rMSE for each coordinate of $\widehat{\beta}^m$. In the second part of Table 1, we report the vector rMSE and MND based on $\hat{\beta}^m$, and the results suggest that our method performs well.

Results Varying N

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

[Table 2 about here.]

Table 2 provides numerical evidence that a larger N helps with overall performance. The sum of absolute bias decreases from 0.0606 to 0.0042 when N increases from $1,000$ to $10,000$. The average size of the estimated sets, rMSE, and MND follow a similar pattern. Notably, even with a relatively small $N = 1,000$, the results remain informative and reasonably accurate, with the rMSE and MND equal to 0.1369 and 0.1159, respectively. We note that T is set to 2 here, which is the minimum required for our method to work. Since our method can extract information from each of the $T(T-1)$ ordered pairs of time periods, a larger T would generally improve the performance of our estimators. Appendix F presents additional simulation results for a larger T.

Finally, we numerically investigate the speed of convergence when we increase N from $1,000$ to $4,000$ and $10,000$ in the second part of Table 2. Compared with the case of $N_0 = 1,000$, the relative ratios of rMSE are 1.84 for $N = 4,000$ and 2.33 for $N = 10,000$, both of which lie between $(N/N_0)^{1/3}$ and $(N/N_0)^{1/2}$. A similar pattern is also observed for calculations based on MND. These results suggest that our estimator converges at a rate slower than $N^{-1/2}$ but faster than $N^{-1/3}$.

6 Empirical Application

6.1 Data and Methodology

We now present an empirical application of the panel multinomial choice model and our proposed estimation method, using NielsenIQ Retail Scanner Data on popcorn sales to examine the effects of display promotions. The data contain weekly store-level information on prices, sales, and display promotion status, collected from approximately 35,000 participating retail stores with point-of-sale systems across the United States.

We focus on popcorn among the wide array of products for two reasons. First, popcorn purchases are more likely to be impulsive, with limited intertemporal planning. Second, popcorn exhibits substantial variation in in-store display promotions, allowing us to estimate how special displays influence consumers' purchase decisions.

We aggregate store-level observations to the designated market area (DMA) level ($N = 205$) for 2015. We focus on the top three brands by market share, pool the remaining brands into a fourth product ("all other products"), and include an outside option of "no purchase." We compute market shares as the dependent variable for each of the $J = 5$ alternatives—the three leading brands, the "all other products" category, and the outside option. The observed product characteristics include price, display-promotion status, and their interaction.[18] Notationally, c denotes each of the $N = 205$ DMAs, j represents each of the $J = 5$ brands, and t indexes the $T = 52$ weeks in 2015. The summary statistics of these variables are provided in Table 3.

[Table 3 about here.]

Since the data are at the DMA level, while our approach was originally developed for individual-level data, we now describe how to adapt the method to the DMA-level set-

[18] We define Price_{cjt} as the weighted-average unit price of all UPCs of the brand j in DMA c during week t. The dataset includes two promotion indicators: display and feature. Given their similarity, we construct Promo_{cjt} as $(\text{feature} \vee \text{display})_{cjt}$. The interaction term $\text{Price}_{cjt} \times \text{Promo}_{cjt}$ is included in X to allow price sensitivity (elasticity) to vary under promotion.

ting. We treat the observed DMA-level market shares s_{cjt} as noisy measurements[19] of $\mathbb{E}\left[y_{cjt}\middle|\mathbf{X}_{ct},\mathbf{A}_c\right]$, i.e.,

$$s_{cjt}=\mathbb{E}\left[y_{cjt}\middle|\mathbf{X}_{ct},\mathbf{A}_c\right]+u_{cjt},\quad \text{with } \mathbb{E}\left[u_c\middle|\mathbf{X}_c,\mathbf{A}_c\right]=0.$$

Then, we use s_{cjt} to nonparametrically estimate the following intertemporal difference:

$$\mathbb{E}\left[s_{cjt}-s_{cjs}\middle|\mathbf{X}_{c,ts}\right]=\int\left(\mathbb{E}\left[y_{cjt}\middle|\mathbf{X}_{ct},\mathbf{A}_c\right]-\mathbb{E}\left[y_{cjs}\middle|\mathbf{X}_{cs},\mathbf{A}_c\right]\right)d\mathbb{P}\left(\mathbf{A}_c\middle|\mathbf{X}_{c,ts}\right).$$

Specifically, we nonparametrically regress $(s_{cjt}-s_{cjs})$ on the second-order polynomial basis functions of $\mathbf{X}_{c,ts}$ with ℓ_1-regularization and 10-fold cross-validation to obtain an estimator $\hat{\gamma}_j$ of $\gamma_j\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right):=\mathbb{E}\left[s_{cjt}-s_{cjs}\middle|\mathbf{X}_{c,ts}=\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)\right]$. Finally, we plug $\hat{\gamma}$ into our second-stage algorithm and compute the (approximate) argmin set $\hat{B}_{\hat{c}}$.

6.2 Results and Discussion

We report our estimation results in Table 4. $\hat{\beta}^m_{\hat{c}}:=\frac{1}{2}\left(\hat{\beta}^l_{\hat{c}}+\hat{\beta}^u_{\hat{c}}\right)$ corresponds to the middle point of the (approximate) argmin set $\hat{B}_{\hat{c}}$ using our method. We show both the exact argmin set ($\hat{c}=0$) and the approximate argmin set with $\hat{c}=0.1\times N^{-\frac{1}{4}}\log(N)\approx 0.14$ for $N=205$. The estimated coefficients for Price (negative) and Promo (positive) are economically intuitive.

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

[Table 4 about here.]

Furthermore, we compare our $\hat{\beta}^m$ with the estimates obtained through four other meth-

[19] Alternatively, with market-level data, one could treat the observed s_{cjt} as a sufficiently good approximation of $\mathbb{E}\left[y_{cjt}\middle|\mathbf{X}_{ct},\mathbf{A}_c\right]$ as in Section 6.1 of Shi, Shum, and Song (2018), in which case our first-stage nonparametric regression is *no longer* required. We did not pursue this approach for two reasons: First, we do not wish to impose the assumption that the observed market shares are measured with negligible errors. Second, we intend this empirical exercise as an illustration of our two-stage procedure, and thus focus on a setting where the first-stage nonparametric regression is required.

ods, i.e., Cyclic Monotonicity (CM) based on Shi, Shum, and Song (2018)[20], OLS, OLS with scalar-valued fixed effects (OLS-FE), the multinomial logit with fixed effects (MLogit-FE), and the random coefficients logit model (RCLM)[21]. Results (normalized to $\mathbb{S}^{D-1}$) are summarized in Table 4.

The OLS estimator for Price is a positive 0.0240, which is counterintuitive. Moreover, displaying the product in the front rows of the store likely makes consumers less price-sensitive, suggesting a positive coefficient on Price×Promo. However, the estimated coefficients for the interaction term using OLS, OLS-FE, and MLogit-FE are all negative. Next, the CM-based estimator for the coefficient of Promo is negative at -0.0567, whereas the estimated coefficient on Price × Promo is a large positive 0.9240. While the aggregate effect of Promo is likely to be positive for most prices observed in the data, it makes the coefficient of Price positive for those promoted products (i.e., $\widehat{\beta}_{Price} + \text{Promo} \times \widehat{\beta}_{Price \times Promo} > 0$ when Promo = 1). Finally, the estimates from RCLM have the same sign as our method. Nonetheless, it reports a smaller estimated coefficient for the interaction term $\text{Price}_{cjt} \times \text{Promo}_{cjt}$, making the effect from Promo on alleviating price sensitivity less significant.

We view the contrast between our findings and those from alternative methods as empirical evidence that, by accommodating more flexible forms of unobserved heterogeneity—via high-dimensional fixed effects that enter consumers' utility functions in an additively non-separable manner—our approach yields more economically plausible results.

6.3 A Possible Explanation via Monte-Carlo Simulations

In this subsection, we provide a possible explanation for the empirical findings reported in Table 4 through simulation analysis. Recall that "Promo" indicates whether a product receives increased in-store exposure by being highlighted by the store. We argue that the negative estimates on $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ reported for traditional methods in Table 4 likely arise from a positive correlation between display promotions and an unobserved index of

[20] We use 2-week cycles for all available weeks in the data for the CM method.
[21] See Appendix F.4 for the details of the RCLM estimator.

price sensitivity.

Specifically, suppose the utility function is

$$u_{ijt} = A_{ij} \times \left(X_{ijt}^{'}\beta_0\right) + \epsilon_{ijt}, \tag{18}$$

where X_{ijt} contains Price, Promo, and Price×Promo, A_{ij} is the ij-specific fixed effect which may capture index sensitivity (which can be thought as inversely related to unobserved brand loyalty), and ϵ_{ijt} is the exogenous random shock. Suppose A_{ij} and Promo_{ijt} are positively correlated, which is reasonable because marketing managers with their expertise are more likely to promote products to which consumers are more price- and promotion-sensitive. Thus, traditional estimation methods based on linearity would be unable to detect such a pattern and wrongly attribute the effect on price elasticities from A_{ij} to Promo.

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

We run 1,000 simulations for each of the five methods in Table 4 to estimate β_0. To replicate the data structure of the empirical exercise, we set $N = 205$, $D = 3$, $J = 4$, and $T = 10$. We report in Table 5 the percentage of simulations that the corresponding method produces correct signs for all coordinates of X_{ijt}.

[Table 5 about here.]

The percentages of simulations where our proposed method produces correct signs for all coordinates of X_{ijt} for $\alpha = 0.15$, 0.3, and 0.5 are 91.50%, 85.90%, and 74.20%, respectively. The accuracy of the estimator is negatively affected by the correlation between $X_{ijt,2}$ (Promo) and A_{ij} (multiplicative fixed effect). In contrast, none of the other methods in Table 5 generates correct signs as ours does. The alternative models, owing to their addi-

tively separable structure,[22] may overlook the positive dependence between Promo and the multiplicative fixed effect A_{ij}, which can bias the resulting estimates.[23]

Intuitively, since products with larger A_{ij} are more likely to be promoted ($X_{ijt,2} = 1$) by the selection of marketing managers, the average effective price sensitivity of promoted products tends to be greater in magnitude than that of non-promoted products. This drives those estimators that ignore such selection effects to produce a negative coefficient on the interaction term. In contrast, our method handles such *non-additive* dependence between observable characteristics and unobserved fixed effects well, illustrating its robustness in these models.

7 Conclusion

This paper develops a method for semiparametric identification and estimation in panel multinomial choice models that feature infinite-dimensional fixed effects and nonadditive utility, thereby accommodating rich forms of unobserved heterogeneity. We also introduce a general identification strategy based on multivariate monotonicity of parametric indices, applicable to a broad class of econometric models defined by the MISC conditions. In addition, we present a computational algorithm that leverages angle-space reparameterization and adaptive-grid search, which prove effective given the nonstandard criterion function implied by our identifying restrictions.

Future research could investigate how our approach might be applied and adapted to other specific microeconometric models within the MISC framework. Along this line, Gao, Li, and Xu (2023)—a companion paper to the present study—illustrates how the approach proposed here can be adapted to the context of dyadic network formation models. Gao

[22] We note that the CM method requires A_{ij} entering the utility function linearly, which is violated in (18).

[23] Notably, RCLM produces "wrong signs" in this simulation exercise, even though it yields the expected signs in the empirical application in Section 6.2. A plausible interpretation is that, while RCLM is more flexible than, for example, MLogit-FE, the unknown selection effect in this dataset may be insufficiently strong to cause RCLM to fail, yet strong enough for MLogit-FE and related methods to do so. This observation highlights the potential value of our method as a robustness-check tool.

and Wang (2025) propose a method for addressing endogeneity in discrete choice models under an adapted time-homogeneity condition. In ongoing work, we are investigating how to combine techniques developed in this line of work to analyze strategic network formation models with endogenous covariates.

References

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

Bach, F. (2017): "Breaking the curse of dimensionality with convex neural networks," *The Journal of Machine Learning Research*, 18, 629–681.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Berry, S. and A. Pakes (2007): "The pure characteristics demand model," *International Economic Review*, 48, 1193–1225.

Berry, S. T. and P. A. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

Blundell, R. and J. L. Powell (2007): "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141, 65–83.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71, 655–679.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

——— (2013): "Penalized sieve estimation and inference of seminonparametric dynamic models: A selective review," in *Advances in Economics and Econometrics: Tenth World*

Congress, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge: Cambridge University Press, Econometric Society Monographs, 485–544.

Chen, X. and H. White (1999): "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., I. Fernández-Val, and W. K. Newey (2019): "Nonseparable multinomial choice models in cross-section and panel data," *Journal of econometrics*, 211, 104–116.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284.

Compiani, G. (2022): "Market counterfactuals and the specification of multiproduct demand: A nonparametric approach," *Quantitative Economics*, 13, 545–591.

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 (2023): "Logical differencing in dyadic network formation models with nontransferable utilities," *Journal of Econometrics*, 235, 302–324.

Gao, W. Y. and R. Wang (2025): "Identification of nonlinear dynamic panels under partial stationarity," *arXiv preprint arXiv:2401.00264*.

Graham, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

Han, A. K. (1987): "Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator," *Journal of Econometrics*, 35, 303–316.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 1251–1271.

Honoré, B. E. and E. Kyriazidou (2000): "Panel data discrete choice models with lagged dependent variables," *Econometrica*, 68, 839–874.

Honoré, B. E. and A. Lewbel (2002): "Semiparametric binary choice panel data models without strictly exogeneous regressors," *Econometrica*, 70, 2053–2063.

Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 505–531.

Khan, S., F. Ouyang, and E. Tamer (2021): "Inference on semiparametric multinomial response models," *Quantitative Economics*, 12, 743–777.

Li, M. (2024): "Identification and estimation in a time-varying endogenous random coefficient panel data model," *arXiv preprint:2110.00982*.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205–228.

——— (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 313–333.

——— (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55, 357–362.

Manski, C. F. and T. S. Thompson (1986): "Operational characteristics of maximum score estimation," *Journal of Econometrics*, 32, 85–108.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*.

Narayanan, S., R. Desiraju, and P. K. Chintagunta (2004): "Return on investment implications for pharmaceutical promotional expenditures: The role of marketing-mix interactions," *Journal of Marketing*, 68, 90–105.

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.

Ota, Y. and T. Otsu (2025): "Specification testing for binary choice model via maximum score," *Working Paper*.

Pakes, A. and J. Porter (2024): "Moment inequalities for multinomial choice with fixed effects," *Quantitative Economics*, 15, 1–25.

Shi, X., M. Shum, and W. Song (2018): "Estimating semi-parametric panel multinomial choice models using cyclic monotonicity," *Econometrica*, 86, 737–761.

Simon, J. L. and J. Arndt (1980): "The shape of the advertising response function," *Journal of Advertising Research*.

Su, L. and S. Jin (2012): "Sieve estimation of panel data models with cross section dependence," *Journal of Econometrics*, 169, 34–47.

Train, K. E. (2009): *Discrete choice methods with simulation*, Cambridge University Press.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak convergence and empirical processes*, Springer.

Vuong, Q. H. (1989): "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, 307–333.

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

Yan, J. and H. I. Yoo (2019): "Semiparametric estimation of the random utility model with rank-ordered choice data," *Journal of Econometrics*, 211, 414–438.

Online Supplemental Material for:

Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects

A Proof of Theorem 2

Proof. Let $\beta^* \in B_0 \backslash \{\beta_0\}$. In the following, we condition on $\mathbf{x}_{st}$ and suppress $\mathbf{x}_{st}$ for notational simplicity. Write $p_{j(t)} := \mathbb{P}(y_{ijt} = 1 | \mathbf{x}_{st})$. Let $\mathcal{J}_+^* := \left\{ j \in \mathcal{J} : (x_s - x_t)' \beta^* \geq 0 \right\}$ and $\mathcal{J}_-^* := \mathcal{J} \backslash \mathcal{J}_+^*$. Then, we have

$$\sum_{j \in \mathcal{J}_+} p_{j(s)} \geq \sum_{j \in \mathcal{J}_+} p_{j(t)}. \tag{19}$$

Since $\sum_j p_{j(s)} = \sum_j p_{j(t)} = 1$, we also have

$$\sum_{j \notin \mathcal{J}_+} p_{j(s)} \leq \sum_{j \notin \mathcal{J}_+} p_{j(t)}. \tag{20}$$

Without loss of generality, relabel products so that

$$\mathcal{J}_+^* = \{J, \ldots, j^*\}, \quad \mathcal{J}_-^* = \{j^* - 1, \ldots, 1\},$$

and

$$p_{J(s)} - p_{J(t)} \geq \ldots \geq p_{j^*(s)} - p_{j^*(t)}, \tag{21}$$

and

$$p_{j^*-1,(s)} - p_{j^*-1,(t)} \geq \ldots \geq p_{1,(s)} - p_{1,(t)}. \tag{22}$$

In words, products are relabeled according to the following lexicographic ascending order:

- Products in $\mathcal{J}_+$ receives larger labels than products in $\mathcal{J}_-$;
- Within each of $\mathcal{J}_+$ and $\mathcal{J}_-$, products are further sorted so that $p_{js} - p_{jt}$ is ascending in (new) product label j.

Given (19) and (21), we must have

$$\sum_{j=h}^{J}\left(p_{j(s)}-p_{j(t)}\right)\geq 0 \quad \forall h\geq j^{*}. \tag{23}$$

In the meanwhile, given (20) and (22), we also have

$$\sum_{j=1}^{h}\left(p_{j(s)}-p_{j(t)}\right)\leq 0 \quad \forall h\leq j^{*}-1.$$

Again, since $\sum_j p_{j(s)}=\sum_j p_{j(t)}=1$, the above analysis implies that

$$\left(1-\sum_{j=1}^{h}p_{j(s)}\right)-\left(1-\sum_{j=1}^{h}p_{j(t)}\right)\geq 0 \quad \forall h\leq j^{*}-1,$$

which is equivalent to

$$\sum_{j=h+1}^{J}\left(p_{j(s)}-p_{j(t)}\right)\geq 0, \quad \forall h\leq j^{*}-1,$$

which is further equivalent to

$$\sum_{j=h}^{J}\left(p_{j(s)}-p_{j(t)}\right)\geq 0, \quad \forall h=1,\ldots,j^{*}. \tag{24}$$

Combining (23) and (24) gives

$$\sum_{j=h}^{J}\left(p_{j(s)}-p_{j(t)}\right)\geq 0 \quad \forall h=1,\ldots,J. \tag{25}$$

Following Pakes and Porter (2024), define the choice mapping $y\left(\mathbf{x}_s,a,\epsilon_s,\beta\right)$ to the product label that is chosen according to model (1) at period s with $(\mathbf{x}_s,a,\epsilon_s,\beta)$, i.e.,

$$y\left(\mathbf{x}_s,a,\epsilon_s,\beta\right)=j \quad \Leftrightarrow \quad u\left(x_{js}^{'}\beta,a_j,\epsilon_{js}\right)>\max_{k\neq j}u\left(x_{ks}^{'}\beta,a_k,\epsilon_{ks}\right).$$

Take a to be any in-support value in $\mathcal{A}$. Define

$$R_{j;s}^{*}:=\left\{\tilde{\epsilon}_s : y\left(\mathbf{x}_s,a,\tilde{\epsilon}_s,\beta^{*}\right)=j\right\}$$

to be the values of ϵ_s^{*} that lead to product j being chosen under $(\mathbf{x}_s,a,\beta^{*})$ in period s.

Recall that $\left(x_{js}-x_{jt}\right)^{'}\beta^{*}\geq 0$ for $j\in\mathcal{J}_{+}^{*}$ and $\left(x_{js}-x_{jt}\right)^{'}\beta^{*}<0$ for $j\in\mathcal{J}\backslash\mathcal{J}_{+}^{*}$. Hence, by the monotonicity of u in its first argument,

$$\bigcup_{j\in\mathcal{J}_{+}^{*}}R_{j;t}^{*}\subseteq\bigcup_{j\in\mathcal{J}_{+}^{*}}R_{j;s}^{*}, \tag{26}$$

and hence

$$R_{j;s}^{*}\cap R_{k;t}^{*}=\emptyset \quad \forall j\in\mathcal{J}_{-}^{*},\ k\in\mathcal{J}_{+}^{*}.$$

Let $R^*_{j,k} := R^*_{j;s} \cap R^*_{k;t}$ and

$$R^*_{j,k,h,l} := R^*_{j,k} \times R^*_{h,l} := \left\{(\tilde{\epsilon}_s, \tilde{\epsilon}_t) : \tilde{\epsilon}_s \in R^*_{j,k}, \tilde{\epsilon}_t \in R^*_{h,l}\right\}.$$

Whenever $R^*_{j,k} \neq \emptyset$, pick any single point $r^*_{j,k} \in R^*_{j,k}$.

We specify the joint distribution of $(\epsilon^*_s, \epsilon^*_t)$ as a discrete distribution over points $\left(r^*_{j,k}, r^*_{h,l}\right)$. We write

$$q^*_{j,k,h,l} := \mathbb{P}\left((\epsilon^*_s, \epsilon^*_t) = \left(r^*_{j,k}, r^*_{h,l}\right) \middle| \mathbf{x}_s, \mathbf{x}_t\right) \equiv \mathbb{P}\left((\epsilon^*_s, \epsilon^*_t) \in R^*_{j,k,h,l} \middle| \mathbf{x}_s, \mathbf{x}_t\right).$$

Hence, a vector $\overline{q}^* = \left(q^*_{j,k,h,l}\right)$ in the unit simplex defines a joint distribution of $(\epsilon^*_s, \epsilon^*_t)$.

By (26), we have

$$q^*_{j,k,h,l} = 0, \ \forall (j,k) \ \text{ or } \ (h,l) \in \mathcal{J}^*_- \times \mathcal{J}^*_+. \tag{27}$$

We set

$$q^*_{j,k,h,l} = 0, \ \forall j < k \text{ or } h < l. \tag{28}$$

Note that $(j,k) \in \mathcal{J}^*_- \times \mathcal{J}^*_+$ implies $j < k$ by the relabeling. Hence, (28) sets a larger class of q_{jkhl} to be zero than those in (27).

Given (28), to specify a joint distribution of $(\epsilon^*_s, \epsilon^*_t)$, we only need to specify

$$q^* := \left(q^*_{j,k,h,l}\right)_{j \geq k, h \geq l}, \tag{29}$$

so that $\overline{q}^* := \left(q^*, \underline{q}^*\right)$ with

$$\underline{q}^* := \left(q^*_{j,k,h,l}\right)_{j<k \text{ or } h<l} = \mathbf{0}.$$

First, q^* needs to satisfy a homogeneity condition. Note that, given (28),

$$\mathbb{P}\left(\epsilon^*_s = r_{jk} \middle| \mathbf{x}_s, \mathbf{x}_t\right) = \sum_{h,l} q^*_{j,k,h,l} = \sum_{h \geq l} q^*_{j,k,h,l}$$

$$\mathbb{P}\left(\epsilon^*_t = r_{jk} \middle| \mathbf{x}_s, \mathbf{x}_t\right) = \sum_{h,l} q^*_{h,l,j,k} = \sum_{h \geq l} q^*_{h,l,j,k}$$

For $j < k$, conditional homogeneity is trivially satisfied, since

$$\mathbb{P}\left(\epsilon^*_s = r_{jk} \middle| \mathbf{x}_s, \mathbf{x}_t\right) = \mathbb{P}\left(\epsilon^*_t = r_{jk} \middle| \mathbf{x}_s, \mathbf{x}_t\right) = 0 \ \forall j < k.$$

Hence, to satisfy homogeneity condition, we just need to ensure

$$\sum_{h\geq l} q^*_{j,k,h,l} = \sum_{h\geq l} q^*_{h,l,j,k}, \quad \forall j \geq k. \tag{30}$$

Second, q^* needs to produce choice probabilities that match the true probabilities. Let

$$\begin{aligned} p^*_{jk} &:= \mathbb{P}\left(y\left(\mathbf{x}_s, a, \epsilon^*_s, \beta\right) = j, y\left(\mathbf{x}_t, a, \epsilon^*_t, \beta\right) = k \middle| \mathbf{x}_s, \mathbf{x}_t\right) \\ &= \mathbb{P}\left((\epsilon^*_s, \epsilon^*_t) \in R^*_{j;s} \times R^*_{k;t} \middle| \mathbf{x}_s, \mathbf{x}_t\right). \end{aligned}$$

Then, the joint choice probabilities

$$p^*_{jk} = \sum_{h=1}^{J}\sum_{l=1}^{J} q^*_{j,h,l,k} = \sum_{h\leq j}\sum_{l\geq k} q^*_{j,h,l,k}$$

To match with true probabilities, we need

$$\sum_{h\leq j}\sum_{l\geq k} q^*_{j,h,l,k} = p_{jk}, \forall j,k = 1,\ldots,J \tag{31}$$

Given $p_{j(s)} = \sum_k p_{jk}$ and $p_{j(t)} = \sum_k p_{kj}$, we can now translate condition (25) about $p_{j(s)} - p_{j(t)}$ into a restriction about (p_{jk}):

$$\sum_{j=h}^{J}\sum_{k} p_{jk} \geq \sum_{j=h}^{J}\sum_{k} p_{kj}, \quad \forall h = 1,\ldots,J. \tag{32}$$

To summarize, given (25), a condition on p_{jk} that we know to hold, we need to show that there exists a $q^* := \left(q^*_{j,k,h,l}\right)_{j\geq k, h\geq l}$ such that (30) and (31) both hold.

The above, however, is exactly the same as the one solved in Pakes and Porter (2024), as summarized by the following lemma. □

Lemma 1 (Nonnegative Solvability of Linear System in Pakes and Porter (2024))**.** *Let* $q^* \equiv \left(q^*_{jkhl}\right)$ *be a* $\left(\frac{1}{2}J\left(J+1\right)\right)^2$*-dimensional vector indexed by* $j,k,h,l \in \{1,\ldots,J\}$ *such that*

$$j \geq k \quad \textit{and} \quad h \geq l.$$

Let $p \equiv (p_{jk})$ *be a* J^2*-dimensional vector indexed by* $j,k \in \{1,\ldots,J\}$ *s.t.* $p \in \Delta^{J^2-1}$*, i.e., p is a (discrete) probability distribution over* J^2 *points defined by* jk*.*

Suppose that

$$\sum_{j=h}^{J}\sum_{k} p_{jk} \geq \sum_{j=h}^{J}\sum_{k} p_{kj}, \quad \forall h = 1,\ldots,J. \tag{33}$$

Then, there exists a nonnegative $q^* \geq 0$ that solves the following joint system of equations:

$$\sum_{h \leq j} \sum_{l \geq k} q^*_{j,h,l,k} = p_{jk}, \quad \forall j,k \in \{1,\ldots,J\}. \tag{34}$$

$$\sum_{h} \sum_{l \leq h} \left(q^*_{h,l,j,k} - q^*_{j,k,h,l} \right) = 0, \quad \forall j,k \in \{1,\ldots,J\} \ \ s.t. \ j \geq k. \tag{35}$$

Lemma 1 above summarizes the part of Pakes and Porter (2024)'s proof of their Theorem 2 that is useful to our setup. Specifically, the inequality constraints in (33) are the same as those in equation (S5) of Pakes and Porter (2024); equations in (34) are the same as those in (S3) of Pakes and Porter (2024); equations in (35) are the same as those in (S4) of Pakes and Porter (2024). The proof in Pakes and Porter (2024) after their equation (S5) then establishes the conclusion of Lemma 1.

B Proof of Theorem 3

We first prove two lemmas before formally proving Theorem 3.

Lemma 2. *$Q : \mathbb{S}^{d-1} \to \mathbb{R}_+$ is continuous.*

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

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

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

Lemma 3. *Under Assumptions 1, 5, and 6,* $\sup_{\beta\in\mathbb{S}^{d-1}} \left|\hat{Q}\left(\beta\right) - Q\left(\beta\right)\right| = O_p\left(c_N\right)$.

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

$$Q_{\tilde{\mathcal{J}},t,s}(\beta,\tilde{\gamma}) := \mathbb{E}\left[G\left(\tilde{\gamma}_{\tilde{\mathcal{J}},t,s}(\mathbf{X}_{i,ts})\right)\lambda_{\tilde{\mathcal{J}}}(\mathbf{X}_{i,ts};\beta)\right],$$
$$\hat{Q}_{\tilde{\mathcal{J}},t,s}(\beta,\tilde{\gamma}) := \frac{1}{n}\sum_{i=1}^{n} G\left(\tilde{\gamma}_{\tilde{\mathcal{J}},t,s}(\mathbf{X}_{i,ts})\right)\lambda_{\tilde{\mathcal{J}}}(\mathbf{X}_{i,ts};\beta),$$

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

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

$$\sup_{\beta\in\mathbb{S}^{d-1}}\sup_{\tilde{\gamma}\in\Gamma}\left|\hat{Q}(\beta,\tilde{\gamma}) - Q(\beta,\tilde{\gamma})\right| = O_p\left(N^{-\frac{1}{2}}\right). \tag{36}$$

Next, by Assumption 4, we have

$$\begin{aligned}\sup_{\beta\in\mathbb{S}^{d-1}}|Q(\beta,\hat{\gamma}) - Q(\beta,\gamma)| &\leq \sup_{\beta\in\mathbb{S}^{d-1}}\int|G(\hat{\gamma}(\mathbf{x}_{ts})) - G(\gamma(\mathbf{x}_{ts}))|\,\lambda_j(\mathbf{x}_{ts};\beta)\,d\mathbb{P}(\mathbf{x}_{ts}) \\ &\leq M\sqrt{\int(\hat{\gamma}(\mathbf{x}_{ts})-\gamma(\mathbf{x}_{ts}))^2\,d\mathbb{P}(\mathbf{x}_{ts})} = O_p(c_N)\end{aligned} \tag{37}$$

by Lipschitz continuity of G, $|\lambda_j|\leq 1$, and the Cauchy-Schwarz inequality. Combining (36) and (37), we have

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

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

Main Proof of Theorem 3

Proof. We verify Condition C.1 in Chernozhukov, Hong, and Tamer (2007, CHT thereafter) so as to apply their Theorem 3.1. Condition C.1(a) on the non-emptiness and compactness of parameter space is satisfied given Theorem 1. Condition C.1(b) on the continuity of the population criterion function Q is proved by Lemma 2. Condition C.1(c) on measurability of the sample criterion function is satisfied by construction. Conditions C.1(d)(e) regarding the uniform convergence of Q_n are satisfied by Lemma 3. Hence, Theorem 3.1.(1) in CHT implies the Hausdorff consistency of $\hat{B}$. The consistency of the point estimator under the additional assumption of point identification (i.e., B_0 is a singleton) follows from Theorem 3.2 of CHT. □

C Sufficient Conditions for Point Identification

In this section, we prove sufficient conditions for the point identification of β_0. For simplicity of notation, we fix $T = 2$, and focus on the conditions that would establish point identification using identifying restrictions from singleton products $j = 1, \ldots, J$. Consequently, these sufficient conditions should be regarded as being strictly more than necessary.

Define $\boldsymbol{\delta}_t := (\delta_{1t}, \ldots, \delta_{Jt})'$, where $\delta_{jt} := x_{jt}'\beta_0$, and recall that $\delta_{ijt} = X_{ijt}'\beta_0$. Denote

$$\psi_j(\boldsymbol{\delta}_t; \mathbf{x}_{ts}, \mathbf{A}_i) := \int \mathbb{1}\left\{u(\delta_{jt}, A_{ij}, \epsilon_{ijt}) > \max_{k\neq j} u(\delta_{kt}, A_{ik}, \epsilon_{ikt})\right\} \mathrm{d}\mathbb{P}(\boldsymbol{\epsilon}_{it} | \mathbf{X}_{i,ts} = \mathbf{x}_{ts}, \mathbf{A}_i)$$

We first impose a strict multivariate monotonicity condition on $\psi_j(\cdot)$.

Assumption 7 (Strict Monotonicity of ψ_j)**.** *For any realized $\mathbf{A}_i$ and $\mathbf{x}_{ts}$, the function $\psi_j(\boldsymbol{\delta}_t; \mathbf{x}_{ts}, \mathbf{A}_i) : \mathbb{R}^J \to \mathbb{R}$ is strictly increasing in δ_{jt} and decreasing in δ_{kt} for $k \neq j$, i.e., if for any two periods t and s it is true that $\delta_{jt} > \delta_{js}$ and $\delta_{kt} < \delta_{ks}$ for all $k \neq j$, then $\psi_j(\boldsymbol{\delta}_t; \mathbf{x}_{ts}, \mathbf{A}_i) > \psi_j(\boldsymbol{\delta}_s; \mathbf{x}_{ts}, \mathbf{A}_i)$.*

We note that Assumption 7 is implied by a stronger version of Assumption 2 together with an additional condition on the support of u given $(\mathbf{X}_i, \mathbf{A}_i)$.

Assumption 2′ (Strict Monotonicity of u)**.** *$u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$ is strictly increasing in the index δ_{ijt}, for every realization of (A_{ij}, ϵ_{ijt}).*

Assumption 2″ (Overlapping Supports)**.** *Conditional on any realization of $\mathbf{X}_i$ and $\mathbf{A}_i$, we have $\bigcap_{j=1}^{J} int\left(Supp\left(u\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right)\right)\right) \neq \emptyset$.*

In particular, Assumption 2″ is directly implied by the assumption of $\text{Supp}\left(u\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right)\right)$ being equal to the real line conditional on any realization of $\mathbf{X}_i$ and $\mathbf{A}_i$, which is again satisfied in additive panel multinomial choice models with scalar fixed effects a la

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

under the assumption of $\text{Supp}\left(\epsilon_{ijt} \middle| \mathbf{X}_i, \mathbf{A}_i\right) = \mathbb{R}$ as commonly imposed in the literature.

Lemma 4. *Assumptions 2′ and 2″ imply Assumption 7.*

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

Assumption 8 (Full-Directional Support of $\Delta\mathbf{X}_i$)**.** *Suppose either (a) or (b) is true:*

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

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

Assumption 8(a) is satisfied when X_{ij} is a continuous random vector. On the other hand, Assumption 8(b) can accommodate discrete regressors generally, but requires one continuous covariate with a large support. Assumption 8 ensures that $\Delta X_{ij}'\beta_0 > 0$ and $\Delta X_{ik}'\beta_0 < 0$ for all $k \neq j$ hold simultaneously with strictly positive probability.

Theorem 4 (Point Identification)**.** *Under Assumptions 1, 3, 7, and 8, β_0 is point identified on $\mathbb{S}^{D-1}$.*

Proof. Fix any $\mathbf{x}_{ts}$ in the support of $\mathbf{X}_i$. By definition of $\gamma_{j,t,s}$, we have

$$\gamma_{j,t,s}(\mathbf{x}_{ts}) = \int \left[\psi_j(\boldsymbol{\delta}_t; \mathbf{x}_{ts}, \mathbf{A}_i) - \psi_j(\boldsymbol{\delta}_s; \mathbf{x}_{ts}, \mathbf{A}_i)\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{i,ts} = \mathbf{x}_{ts}\right).$$

Hence, under Assumption 7, we have

$$\delta_{jt} > \delta_{js} \text{ and } \delta_{kt} < \delta_{ks} \text{ for all } k \neq j \quad \Rightarrow \quad \gamma_{j,t,s}(\mathbf{x}_{ts}) > 0, \tag{38}$$

since $\psi_j(\boldsymbol{\delta}_t; \mathbf{x}_{ts}, \mathbf{A}_i) > \psi_j(\boldsymbol{\delta}_s; \mathbf{x}_{ts}, \mathbf{A}_i)$ for every realization of $\mathbf{A}_i$ given $\mathbf{X}_{i,ts} = \mathbf{x}_{ts}$. Together with Assumption 8, we deduce that

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

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

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

As $\beta \neq \beta_0$, by Assumption 8 we have

$$\mathbb{P}(\Delta\mathbf{X}_i \in H_j(\beta)) > 0. \tag{39}$$

Moreover, for any realization of $\mathbf{X}_i$ such that $\Delta\mathbf{X}_i \in H_j(\beta)$, we must have: (i) $\gamma_{j,t,s}(\mathbf{X}_i) > 0$ by (38), and (ii)

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

so that

$$G(\gamma_j(\mathbf{X}_i))\, \lambda_j(\Delta\mathbf{X}_i, \beta) = G(\gamma_j(\mathbf{X}_i)) > 0$$

for all such $\mathbf{X}_i$. Hence, we have

$$\mathbb{E}\left[G(\gamma_j(\mathbf{X}_i)) \middle| \Delta\mathbf{X}_i \in H_j(\beta)\right] > 0. \tag{40}$$

Combining (39) and (40), we have

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

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

□

D More on the Adaptive-Grid Algorithm

We elaborate on the adaptive-grid algorithm described in Section 3.3 and provide additional details regarding its practical implementation. A graphical illustration of the algorithm is presented in Figure 1.

1. **Initialization and Coarse Grid Search.** The search begins by defining a coarse, uniform grid of M_0^{D-1} points over the initial parameter space Θ. The code sets M_0 via the `M_Step` parameter. The objective function $Q(\beta)$ is evaluated at each point on this grid. This stage then iteratively refines the search space. After each evaluation, the algorithm prunes the parameter space by retaining only the region containing the points that yielded the lowest 20% of objective function values. A new, equally-sized grid is then constructed within this smaller, more promising region, and the process repeats. This iterative pruning continues for a fixed number of loops or until the minimum objective function value stabilizes, effectively narrowing the search to a region likely to contain the global minimum. The procedure includes a check to recenter the search space if the minimum appears to lie on a boundary of the angular coordinates, which prevents erroneously discarding a wrapped-around solution.

2. **Fine Grid Refinement.** Upon completion of the coarse search, the algorithm enters a refinement stage. It constructs a new, finer grid within a small neighborhood surrounding the set of candidate points that produced the minimum objective function value in the previous stage. The resolution of this grid is determined by a smaller step

size, `Tol_Step`. By re-evaluating the objective function on this denser set of points, this stage seeks to improve upon the initial estimate. If a new, lower minimum value is found, the process repeats around this new set of points. This stage concludes when no further improvement in the minimum of the objective function can be achieved, indicating that the algorithm has converged to a local minimum within the refined region.

3. **Boundary Identification.** The final stage aims to precisely delineate the identified set of minimizing parameters. A very fine grid is constructed in the immediate buffer zone *around* the set of optimal points found in the prior stage, while excluding the interior of this set. This allows the algorithm to meticulously probe the boundary of the solution set. Any points on this boundary that also yield the minimum objective function value are added to the solution set. This step repeats, further reducing the grid step size until a pre-specified tolerance (`Tol`) is met and a clear separation exists between the parameters that minimize the function and those that do not. This ensures the final identified set is not only optimal but also well-defined.

[Figure 1 about here.]

Figure 1 provides a graphical illustration of our adaptive-grid algorithm to find the minimizer of the criterion function over the parameter space Θ. The figure corresponds to a re-parameterization in the 2-dimensional angle space of a 3-dimensional unit ball, as specified in equation (14) of the paper for $D = 3$. The pink area represents the initial angle space. The light blue rectangle corresponds to $\underline{\Theta}^{(1)}$ defined in (16). The dark blue rectangle represents the identified set Θ_0, which is the best we can hope for. The disconnectedness of the light blue rectangle corresponds to Step 3 of the algorithm presented in Section 3.3. This is largely a programming nuisance: it means that if $\underline{\Theta}^{(1)} \not\subset \Theta^{(1)}$ due to the special structure of the angle space, we add 2π to $\underline{\Theta}^{(1)}$ such that it is connected and lies within $\Theta^{(1)}$. Note that adding 2π to the first coordinate of θ does not change the value of θ nor β.

E Counterfactual Analysis in Long Panels

So far, we have focused on the identification and estimation of the index parameter β_0. While β_0 itself contains rich information about the unobserved preference of consumers, we are often also interested in counterfactual parameters defined as some functional of not only β_0, but also other unknown components of the model. In this section, we discuss how the estimate $\hat{\beta}$ of β_0, and the computed indices based on $\hat{\beta}$, can be used to estimate more sophisticated counterfactual parameters.

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

$$\mu_j\left(\overline{\mathbf{X}}\right) := \int \mathbb{E}\left[y_{ijt} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{A}_i\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i\right), \tag{41}$$

Demand elasticities may be further computed as $\nabla \mu_j\left(\overline{\mathbf{X}}\right)$, which gives the marginal effect of an exogenous change in certain observable characteristics on consumer choices.[24]

To achieve this separation, we seek to identify and estimate the integrand $\mathbb{E}\left[y_{ijt} | \, \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{A}_i\right]$, which is a function of β_0. Conditional on $\boldsymbol{\delta}_{it} = \overline{\mathbf{X}}'\beta_0 =: \overline{\boldsymbol{\delta}}$ and $\mathbf{A}_i$, by our model specification (1) 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., $T \to \infty$. Suppose the conditions in Appendix C are also satisfied so that δ_0 is point identified. We can identify $\psi_{ij}\left(\overline{\boldsymbol{\delta}}\right)$ for each fixed i and product j and estimate it via nonparametric regression of y_{ijt} on $(\delta_{i1t}, \ldots, \delta_{iJt})$ across $t = 1, \ldots, T$, under suitable

[24] 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 $\mathbb{P}\left(\mathbf{A}_i | \, \mathbf{X}_{it} = \overline{\mathbf{X}}\right)$. 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 $\mu_j\left(\overline{\mathbf{X}}\right)$ 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*.

stationarity and weak dependence conditions on the error terms. See, for example, Su and Jin (2012) for a sieve estimator of functional fixed effects in a long panel setting.

We are now in the position to estimate the counterfactual market shares of product $j = 1, \ldots, J$ at any counterfactual $\overline{\mathbf{X}}$, where $\overline{\mathbf{X}}$ is a $D \times J$ matrix. First, we use the estimated $\hat{\beta}$ to compute the counterfactual index $\hat{\boldsymbol{\delta}}(\cdot)$ evaluated at $\overline{\mathbf{X}}$, i.e., $\hat{\boldsymbol{\delta}}\left(\overline{\mathbf{X}}\right) = \overline{\mathbf{X}}'\hat{\beta}$. Then, we obtain $\hat{\psi}_{ij}\left(\hat{\boldsymbol{\delta}}\left(\overline{\mathbf{X}}\right)\right)$ by plugging $\hat{\boldsymbol{\delta}}\left(\overline{\mathbf{X}}\right)$ into the nonparametric estimate $\hat{\psi}_{ij}(\cdot)$ for each fixed i and product j. Finally, we estimate $\mu_j\left(\overline{\mathbf{X}}\right)$ for product j by averaging over individuals in the sample, i.e., $\mu_j\left(\overline{\mathbf{X}}\right) = \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{ij}\left(\hat{\boldsymbol{\delta}}\left(\overline{\mathbf{X}}\right)\right)$. The last step corresponds to the integration with respect to the probability measure of A in (41).

F Additional Simulation Results

F.1 Adaptive-Grid Computation Algorithm

In this section, we illustrate a typical output of our second-step computation algorithm based on the adaptive-grid search over the angle space, and show that the algorithm works well. For this purpose we consider a simplified DGP without fixed effect A_{ij}. We draw each of $X_{ijt,d}$ independently across each dimension d from the standard normal distribution, and set the distribution of the idiosyncratic shock to be $\epsilon_{ijt} \sim TIEV(0,1)$, so that we can skip the first-step estimation and directly compute the true conditional choice probability. Note that the conditions for point identification of β_0 are satisfied. Because we are only seeking to illustrate the validity of the algorithm itself, we set N to be large with $N = 10^7$ and fix $D = 3,\ J = 3,$ and $T = 2$. Then, we apply our adaptive-grid algorithm to search for β_0.

[Figure 2 about here.]

Figure 2 shows how our computational algorithm works in finding the true unknown θ_0, the angle representation of the true β_0 in the Θ space. The horizontal and vertical axes correspond to the two polar coordinates that are associated with $\mathbb{S}^2$. The blue dots

represent the points that our algorithm searches over but find *not* to be minimizers of the sample criterion $\hat{Q}$. The black box indicates the area that the minimizers for the sample criterion $\hat{Q}$ lie within, or more precisely, a rectangular enclosure of the numerical argmin set. The big black dot stands for the true parameter value $\theta_0 = (0.4205, 0.4636)'$.

It is evident from Figure 2 that our adaptive-grid algorithm is able to correctly locate an area that covers the true θ_0, which lies within the small black box representing the estimated set of $\hat{\theta}$, demonstrating the efficacy of the algorithm. Besides, it is worth mentioning that our algorithm computes reasonably fast, as it first performs a rough search on the whole unit sphere $\mathbb{S}^2$, then focuses on the area where the minimizers are most likely to lie. In the last few rounds of search, the algorithm evaluates the criterion function $\hat{Q}$ on a relatively small area of points shown by those blue and red dots in Figure 2 until the desired level of accuracy is achieved.

For a more transparent representation, we translate the angles θ in the polar coordinates into unit vectors β on the unit sphere $\mathbb{S}^2$ and show it in Figure 3, which is now plotted on $\mathbb{S}^2 \subseteq \mathbb{R}^3$. Again the blue dots represent the points that do not achieve the minimum of $\hat{Q}$; the black box shows an enclosing set of the minimizers of $\hat{Q}$. The big black dot represents the true parameter value β_0, which resides inside the black box of the minimizers of $\hat{Q}$. Figure 3 illustrates that our computation algorithm is able to locate a tight area around β_0.

[Figure 3 about here.]

F.2 Estimation without Point Identification

We now investigate the performance of our estimator when point identification fails. To make things comparable, we fix (N, D, J, T) as in the baseline case, but modify the configuration in two different ways. We maintain the point identification in one setting but lose the point identification in the other. Specifically, we set $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $X_{ijt,1} \sim \mathcal{U}[-1, 1]$, $X_{ijt,2} = Z_i + \mathcal{N}(0, 6)$, and $X_{ijt,3} \sim \mathcal{N}(0, 1)$ for the point identified case. For the DGP without point identification, we let $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $P(X_{ijt,1} = 1/2) = P(X_{ijt,1} = -1/2) = 0.5$,

$X_{ijt,2} = Z_i + \mathcal{U}\left[-\sqrt{6}, \sqrt{6}\right]$, and $P(X_{ijt,3} = 1/2) = P(X_{ijt,3} = -1/2) = 0.5$. The construction of A_{ij} is the same as in the DGP for the baseline results in Table 1. We deliberately control the location and scale of each variable to be comparable across the two configurations, with the only differences being the presence of discreteness and boundedness of supports. When point identification fails, we compute the set estimator $\hat{B}_{\hat{c}}$ of (14) with $\hat{c} > 0$. Table 6 contains simulation results under the two configurations, with different choices of $\hat{c}$ when point identification fails.[25]

[Table 6 about here.]

Across rows in (i) and (ii), we see that the lack of point identification does negatively affect the performance of our estimates, but the impact is limited to a moderate degree. Within rows in (ii), we observe that, as expected, a more conservative choice of the constant $\hat{c}$ worsens performance of the upper and lower bounds by enlarging the estimated sets; meanwhile, it appears that the size and the performance of our estimator based on $\hat{\beta}^m$ is not sensitive to the choice of $\hat{c}$.

F.3 Results Varying D, J, T

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

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

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

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

We report in Table 7 the performance of our estimators for each of the corresponding configurations across all $M = 1,000$ simulations.

[Table 7 about here.]

From Table 7 we find a larger T improves the performance of our estimator, which is arguably more practically relevant given the increasing availability of long panel data. The improvement in performance with larger T is because our method can extract more information from $T \times (T-1)$ ordered pairs of time periods which effectively increase the total number of observations. We also find that increase in D or J adversely affects the performance of our estimator, which is expected because more information is required to estimate more covariates or deal with more alternatives. For example, when J is 3 and T is 4, an increase in the dimension of product characteristics D from 3 to 4 increases the rMSE from 0.0411 to 0.0497. The change in performance for increasing J is more significant. For instance, when $D = 4$ and $T = 4$, an increase in J from 3 to 4 increases the MND from 0.0475 to 0.1003.

F.4 Robustness Against the Blue-Bus/Red-Bus Problem

We now demonstrate the robustness of our method against the Blue-Bus/Red-Bus problem, or more precisely, a case where the product set includes highly substitutable alternatives.

Specifically, we modify the DGP in our baseline simulation by increasing J from 3 to 4, and introduce the fourth product as an indistinguishable duplicate of the third one: they share the same observable and unobservable characteristics, including the observable characteristics X, unobserved fixed effect A, and the unobserved error term ϵ. We equally split the market share between $j = 3$ and $j = 4$.

Using the above DGP, we compare our method's performance to that of the RCLM estimator. For the RCLM estimator, we simulate $M = 1,000$ random coefficients $\beta_i = \beta + \eta_i$ where $\eta_i \sim N(0, I_D)$ for the consumers with heterogeneous preferences, where I_D is a $D \times D$

identity matrix. The predicted market shares are thus computed as

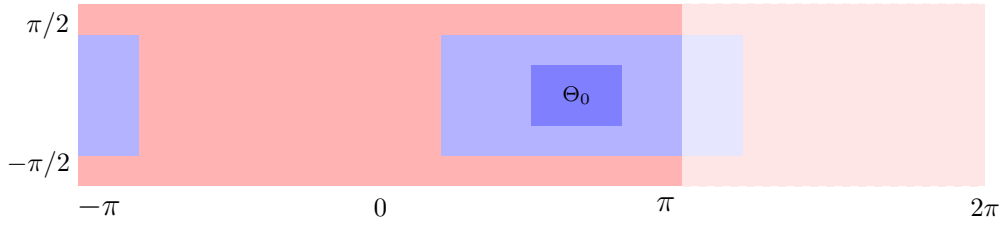
$$\begin{aligned}\sigma_{jt}\left(X_{i\cdot t};\beta\right) &= \mathbb{E}\left[\mathbb{1}\left\{U_{ijt} > \max_{k\neq j} U_{ikt}\right\}\middle| X_{i\cdot t}\right] \\ &= \int \frac{\exp\left(X_{ijt}'\beta + X_{ijt}'\eta_i\right)}{\sum_{k=1}^{J}\exp\left(X_{ikt}'\beta + X_{ikt}'\eta_i\right)} dF_\eta\left(\eta_i\right) \\ &\simeq \frac{1}{M}\sum_{m=1}^{M}\frac{\exp\left(X_{ijt}'\beta + X_{ijt}'\eta_i^{(m)}\right)}{\sum_{k=1}^{J}\exp\left(X_{ikt}'\beta + X_{ikt}'\eta_i^{(m)}\right)}. \end{aligned} \tag{42}$$

For each round of simulation, we compute predicted market shares by numerically integrating over the random taste distribution over M agents. We then choose the candidate β over a fine grid of $L = 10,000$ points from $\mathbb{S}^{D-1}$ that achieves the smallest distance in the Euclidean norm between the predicted and observed market shares as $\hat{\beta}$. We use $N = 1,000$, $D = 3$, $J = 4$, $T = 2$, and B =1,000 for this simulation exercise.

Table 8 summarizes the results. The sum of absolute biases of $\hat{\beta}$ under our method is 0.0452, which is markedly lower than that of the RCLM estimator (0.8689). The average size of the estimated sets is 0.0585 for our method. Since, by construction, the RCLM estimator is a point estimator, its average size is zero across simulations. Finally, the rMSE and MND are significantly lower for our estimator than for the RCLM. The results show that, in the presence of highly similar products (with highly correlated unobserved heterogeneity), the RCLM estimator cannot recover the structural parameters accurately. In contrast, our method is robust to arbitrary dependence structures in the unobserved heterogeneity terms across products, leading to more accurate estimates of the structural parameters.

[Table 8 about here.]

Figure 1: An Adaptive-Grid Algorithm



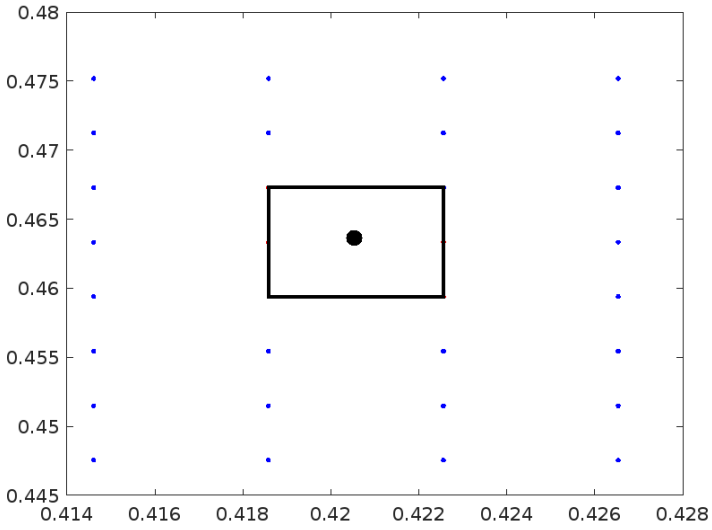


Figure 2: The Argmin Set in Θ

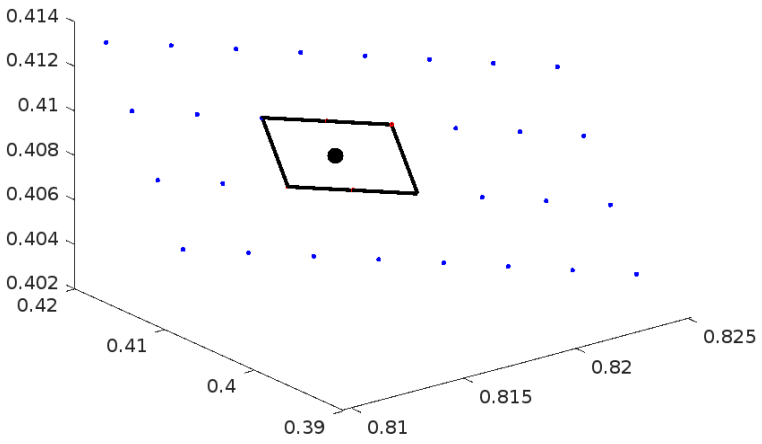


Figure 3: The Argmin Set in $\mathbb{S}^2$

Table 1: Baseline Performance

	$\beta_0 = (0.82, 0.41, 0.41)'$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
mid bias	$\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{m} - \beta_{0,d}\right)$	-0.0005	-0.0003	-0.0034
upper bias	$\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{u} - \beta_{0,d}\right)$	0.0075	0.0080	0.0083
lower bias	$\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{l} - \beta_{0,d}\right)$	-0.0085	-0.0086	-0.0150
mean(u−l)	$\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{u} - \hat{\beta}_{b,d}^{l}\right)$	0.0160	0.0166	0.0233
standard deviation	$\sqrt{\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{m} - \overline{\hat{\beta}_{d}^{m}}\right)^{2}}$	0.0299	0.0308	0.0431
root MSE (by coordinate)	$\left(\frac{1}{M}\sum_{b=1}^{M}\left(\hat{\beta}_{b,d}^{m} - \beta_{0,d}\right)^{2}\right)^{1/2}$	0.0288	0.0296	0.0417
root MSE (whole vector)	$\left(\frac{1}{M}\sum_{b=1}^{M}\left\|\hat{\beta}_{b}^{m} - \beta_{0}\right\|^{2}\right)^{1/2}$		0.0587	
mean norm deviations (MND)	$\frac{1}{M}\sum_{b=1}^{M}\left\|\hat{\beta}_{b}^{m} - \beta_{0}\right\|$		0.0511	

Table 2: Performance under Varying N

	$\sum_d \lvert\text{bias}_d\rvert$	$\sum_d \text{mean(u-l)}_d$	rMSE	MND
$N = 10,000$	0.0042	0.0560	0.0587	0.0511
$N = 4,000$	0.0136	0.0865	0.0742	0.0650
$N = 1,000$	0.0606	0.1664	0.1369	0.1159
	$\left(\frac{N}{1,000}\right)^{1/2}$	$\left(\frac{N}{1,000}\right)^{1/3}$	$\frac{\text{rMSE}_{1000}}{\text{rMSE}_N}$	$\frac{\text{MND}_{1000}}{\text{MND}_N}$
$N = 10,000$	3.16	2.15	2.33	2.27
$N = 4,000$	2.00	1.59	1.84	1.78

Table 3: Empirical Application: Summary Statistics

	mean	s.d.	min	max
DMA-level Market Share s_{cjt}	25.06%	21.59%	0.08%	96.69%
Price_{cjt}	0.4924	0.1803	0.1094	1.3587
Promo_{cjt}	0.0282	0.0377	0.0000	0.5000
$\text{Price}_{cjt} \times \text{Promo}_{cjt}$	0.0136	0.0203	0.0000	0.4505

Table 4: Empirical Illustration: Comparison of Results

	$\hat{\beta}^{m}_{\hat{c}=0}$	$\hat{\beta}^{m}_{\hat{c}=0.14}$	$\hat{\beta}^{CyclicMono}$	$\hat{\beta}^{OLS}$	$\hat{\beta}^{OLS-FE}$	$\hat{\beta}^{MLogit-FE}$	$\hat{\beta}^{RCLM}$
Price_{cjt}	-0.9351	-0.9283	-0.3781	0.0240	-0.3807	-0.6249	-0.9705
Promo_{cjt}	0.1793	0.1912	-0.0567	0.5760	0.5976	0.5881	0.2157
$\text{Price}_{cjt} \times \text{Promo}_{cjt}$	0.2687	0.2505	0.9240	-0.8171	-0.7057	-0.5135	0.1078

Table 5: Percentage of Correct Signs of Estimated Coefficients

α	$\hat{\beta}^{m}$	$\hat{\beta}^{CyclicMono}$	$\hat{\beta}^{OLS}$	$\hat{\beta}^{OLS-FE}$	$\hat{\beta}^{MLogit-FE}$	$\hat{\beta}^{RCLM}$
0.15	91.50%	0.00%	0.00%	0.00%	28.00%	0.00%
0.30	85.90%	0.00%	0.00%	0.00%	0.20%	0.00%
0.50	74.20%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 6: Performance with and without Point ID: Further Examination

point ID ?	$\hat{c}$	rMSE			MND		
		$\hat{\beta}^m$	$\hat{\beta}^u$	$\hat{\beta}^l$	$\hat{\beta}^m$	$\hat{\beta}^u$	$\hat{\beta}^l$
(i) yes	-	0.0574	0.0636	0.0665	0.0501	0.0577	0.0602
(ii) no	0.01	0.0624	0.0657	0.0711	0.0536	0.0590	0.0630
	0.1	0.0521	0.0820	0.0966	0.0444	0.0782	0.0904
	1	0.0476	0.2198	0.2631	0.0458	0.2190	0.2619

Table 7: Performance Varying D, J, T

rMSE	$J=3$		$J=4$	
	$T=2$	$T=4$	$T=2$	$T=4$
$D=3$	0.0594	0.0411	0.1321	0.0982
$D=4$	0.0725	0.0497	0.1384	0.1091
MND	$J=3$		$J=4$	
	$T=2$	$T=4$	$T=2$	$T=4$
$D=3$	0.0522	0.0363	0.1154	0.0860
$D=4$	0.0660	0.0475	0.1264	0.1003

Table 8: Illustration of the "Blue-Bus/Red-Bus" Problem

	$\sum_d \lvert\text{bias}_d\rvert$	$\sum_d \text{mean(u-l)}_d$	rMSE	MND
Current Paper	0.0452	0.0585	0.1389	0.1210
RCLM	0.8689	-	0.6047	0.6042