# Robust Counterfactual Analysis for Nonlinear Panel Data Models*

Yan Liu†

(Job market paper)

October 22, 2024

Please click here for the latest version.

**Abstract**

This paper studies robust counterfactual analysis in a wide variety of nonlinear panel data models. I focus on counterfactual predictions of the behavior of an outcome variable under exogenous manipulations of endogenous explanatory variables. I avoid parametric distributional assumptions and only impose time homogeneity on the distribution of unobserved heterogeneity. I derive the sharp identified set for the distribution of the counterfactual outcome, noting that point identification is impossible in general. I provide tractable implementation procedures for popular nonlinear models, including binary choice, ordered choice, censored regression, and multinomial choice, by exploiting an index separability condition. I propose inference for sharp bounds on counterfactual probabilities based on aggregate intersection bounds and Bonferroni-adjusted confidence intervals. As empirical illustrations, I apply my approach to actual data to predict female labor force participation rates under counterfactual fertility scenarios, as well as market shares of different saltine cracker brands under counterfactual pricing schemes.

*Keywords:* Average structural function, discrete choice, semiparametric, sharp partial identification

*This paper supersedes my earlier paper "Counterfactual Analysis for Semiparametric Panel Multinomial Choice Models". I am indebted to Hiroaki Kaido for his guidance and encouragement throughout this project. For helpful comments and discussions, I thank Rubaiyat Alam, Shuowen Chen, Eli Ben-Michael, Xu Cheng, Iván Fernández-Val, Jean-Jacques Forneron, Stella Hong, Shakeeb Khan, Lixiong Li, Aureo de Paula, Pierre Perron, Kirill Ponomarev, Marc Rysman, Davide Viviano, and participants of the econometrics seminar at Boston University, BU-BC Greenline Workshop in Econometrics 2023, and AMES2024-China. All errors are my own.

†Department of Economics, Boston University. Email: yanliu@bu.edu.

# 1 Introduction

A frequent goal in empirical research is to predict the counterfactual behavior of an outcome variable under *ceteris paribus* manipulations of endogenous explanatory variables. For instance, the policymaker may want to predict the counterfactual probability of a woman participating in the labor force if her fertility and husband's income were externally set at some values. This is an important policy question related to offering and subsidizing child care. It has been common practice in this context to use a threshold-crossing model where the latent index is a function of explanatory variables, including fertility and husband's income, and unobserved heterogeneity. Unobserved heterogeneity enters the outcome equation in a non-additive manner and can depend on latent factors determining fertility and husband's income, such as household productivity and access to job networks. As a result, predicting the counterfactual female labor participation rate requires knowledge of not only index coefficients but also the distribution of unobserved heterogeneity.

Panel data offers the possibility of controlling for unobserved heterogeneity by utilizing multiple observations of a single economic unit over time. This possibility extends to nonlinear models, which naturally arise in the context of discrete outcomes. Since the seminal work of Manski (1987), the literature on semiparametric nonlinear panel data models has developed methods to identify structural parameters, such as index coefficients, which are insufficient for making counterfactual predictions. What has been missing is a framework to systematically quantify what can be learned about the distribution of unobserved heterogeneity. This paper aims to fill this gap.

This paper develops a method for robust counterfactual analysis in nonlinear panel data models. The only restriction imposed on the distribution of unobserved heterogeneity is *time homogeneity*, which can be interpreted as "time is randomly assigned" or "time is an instrument" (Chernozhukov, Fernández-Val, Hahn, and Newey, 2013) and formally justifies combining information from an individual's observations over time. At the same time, this assumption is general enough to allow for flexible dependence between unobserved heterogeneity and explanatory variables. I note that when the outcome distribution exhibits mass points (e.g., discrete or mixed), it is generally impossible to point identify both structural parameters and the distribution of the counterfactual outcome without further assumptions. Therefore, I focus on cases where structural parameters are point-identified and derive the sharp identified set for the distribution of the counterfactual outcome.

The main idea of identification is to collect all values of unobserved heterogeneity for which outcomes are identical into what I refer to as "$U$-level sets." Identified sets of counterfactuals defined through $U$-level sets are guaranteed to be sharp, i.e., they use all available information. The time homogeneity assumption simplifies the sharp identified set as intersections across time periods. Nonetheless, calculating the sharp identified set can still be challenging because it involves searching over all distributions of unobserved heterogeneity. I provide tractable implementation procedures that bypass this search for two important classes of nonlinear models: monotone transformation models, such as binary choice, ordered choice, censored regression, and multinomial choice models. To this end, I exploit an index separability condition that connects the comparison of index functions of explanatory variables under factual and counterfactual scenarios to the set inclusion relationship of $U$-level sets, which can be translated into the comparison of the distributions of observed and counterfactual outcomes. In this way, I generate identifying restrictions on the distribution of the counterfactual outcome directly from observed data. While my baseline framework focuses on static settings, I also consider an extension of my identification strategy to dynamic binary choice models.

When it comes to estimation and inference, I target summary measures of the distribution of the counterfactual outcome in the spirit of the *average structural function* introduced in Blundell and Powell (2003, 2004), which are typically counterfactual probabilities for discrete outcomes. Sharp bounds on counterfactual probabilities take the form of aggregate intersection bounds (cf. Semenova (2024)). Inference poses a challenge due to the uncertainty in structural parameters. I propose a two-step procedure. In the first step, a confidence region for structural parameters is constructed. In the second step, for each value of structural parameters, I provide conditions for the pointwise asymptotic normality of the estimator for aggregate intersection bounds, giving rise to simple pointwise confidence intervals for counterfactual probabilities. I combine these two steps using a Bonferroni adjustment.

As an empirical illustration, I apply my approach to U.S. and U.K. data to predict female labor force participation rates under counterfactual fertility scenarios. The bounds reveal a common pattern in both samples: having one more infant or preschooler decreases labor force participation, while the effect of one more school-age child is ambiguous. I also demonstrate the application of my approach in multinomial choice models to predict market shares of different saltine cracker brands under counterfactual pricing schemes.

This paper contributes to three strands of literature. First, there is a growing literature on semiparametric identification of nonlinear panel data models, including Manski (1987), Honoré and Kyriazidou (2000), Khan, Ponomareva, and Tamer (2016, 2023), Shi, Shum, and Song (2018), Gao and Li (2020), Khan, Ouyang, and Tamer (2021), Botosaru, Muris, and Pendakur (2023), Chesher, Rosen, and Zhang (2024), Gao and Wang (2024), Pakes and Porter (2024). It is well known that structural parameters, such as index coefficients, can be identified under time homogeneity, but little is known about how to identify counterfactuals, which also require the full distribution of unobserved heterogeneity. I take a step forward to bound counterfactuals under these assumptions. The framework of Chesher et al. (2024) potentially permits counterfactual analysis. They impose a fixed effects structure on unobserved heterogeneity while leaving the distribution of fixed effects completely unrestricted. As a result, their approach cannot predict the counterfactual probability in a single period, which is my focus, because fixed effects can be arbitrarily moved to justify any outcome. When specialized to multinomial choice models, set inclusion relationships of $U$-level sets also underlie the identification strategy of Pakes and Porter (2024). They focus on deriving sharp identifying restrictions on structural parameters in the case with only two time periods. In contrast, my object of interest is counterfactuals, and my sharpness results apply to longer panels.

Second, this paper complements the literature on identification of counterfactuals in discrete outcome models, including Manski (2007), Chiong, Hsieh, and Shum (2021), Gu, Russell, and Stringham (2024), Tebaldi, Torgovitsky, and Yang (2023). Manski (2007) focused on counterfactual scenarios concerning unrealized choice sets. Chiong et al. (2021) assumed exogeneity of product-specific attributes and proposed using *cyclic monotonicity* to bound counterfactual market shares under changes in these attributes. Tebaldi et al. (2023) restricted explanatory variables to be finitely supported. In this case, searching over latent distributions reduces to a finite-dimensional problem characterized by a finite partition of the space of unobserved heterogeneity, termed the *minimal relevant partition*. Gu et al. (2024) extended this insight to account for model misspecification and model incompleteness. An obvious feature of my approach is that I exploit the panel data structure. Moreover, I allow explanatory variables to be both endogenous and continuous.

Third, this paper adds to the literature on identification of counterfactuals in nonlinear panel data models, including Hoderlein and White (2012), Chernozhukov et al. (2013), Chernozhukov, Fernández-Val, and Newey (2019), Liu, Poirier, and Shiu (2021), Davezies, D'Haultfoeuille, and Laage (2022), Botosaru and Muris (2024), Pakel and Weidner (2024). The identification results of Hoderlein and White (2012) and Chernozhukov et al. (2019) are confined to the subpopulation of "stayers", i.e., the population for which explanatory variables do not change over time. Chernozhukov et al. (2013) only considered finitely supported explanatory variables. By comparison, I handle counterfactuals that are averaged over the whole population and continuous explanatory variables. Liu et al. (2021) concentrated on binary choice models and achieved point identification of average effects by imposing index sufficiency on the distribution of fixed effects. Davezies et al. (2022) and Pakel and Weidner (2024) did not restrict the distribution of fixed effects but relied on parametric distributional assumptions on idiosyncratic shocks (e.g., fixed effects logit). They provided bounds on average effects. Botosaru and Muris (2024) derived bounds on counterfactual survival probabilities in monotone transformation models. My results differ in that I work with weaker assumptions and cover a relatively wide variety of nonlinear models.

The remainder of this paper is organized as follows. Section 2 outlines the setup and specifies the type of counterfactuals under consideration. Section 3 presents the sharp identified set for the distribution of the counterfactual outcome. Section 4 discusses the tractable implementation of the sharp identified set. Section 5 addresses estimation and inference. Section 6 gives numerical results for the sharp identified set. Section 7 contains empirical illustrations using data on female labor force participation and purchases of saltine crackers. Section 8 explores the extension to dynamic binary choice models. Section 9 concludes. Proofs and simulation results are collected in the Appendix.

## 2 Setup

This paper considers panel data models of the form:

$$Y_{it} = g(X_{it}, U_{it}; \theta_0),\ \ i = 1, \dots, N, t = 1, \dots, T,$$

where $Y_{it} \in \mathcal{Y} \subseteq \mathbb{R}$ denotes an observed scalar outcome, $X_{it} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ denotes explanatory variables, $U_{it} \in \mathbb{R}^{d_u}$ denotes unobserved heterogeneity, and $g$ is a function known up to a finite-dimensional parameter $\theta_0$. Write $X_i = (X_{i1}, \dots, X_{iT})$. Throughout, I assume that the data are independent and identically distributed (i.i.d.) across $i$. For the identification analysis in Sections 3, 4, and 8, I drop the $i$ subscript to simplify the notation.

**Example 1** (Binary choice model)**.** Consider the model

$$Y_{it} = 1\{X_{it}^\top \beta_0 + U_{it} \geq 0\},$$

where $\beta_0 \in \mathbb{R}^{d_x}$ is a vector of unknown coefficients. Here $\theta_0 = \beta_0$ and $\mathcal{Y} = \{0, 1\}$.

**Example 2** (Ordered choice model)**.** Consider the model

$$Y_{it} = \sum_{j=0}^{J} 1\{X_{it}^{\top}\beta_0 + U_{it} \geq \gamma_0^j\},$$

where $\beta_0 \in \mathbb{R}^{d_x}$ is a vector of unknown coefficients, and $\gamma_0 = (\gamma_0^0, \gamma_0^1, \ldots, \gamma_0^J)$ are unknown thresholds satisfying $\gamma_0^j > \gamma_0^{j-1}$ and $\gamma_0^0 = -\infty$. Here $\theta_0 = (\beta_0, \gamma_0)$ and $\mathcal{Y} = \{0, 1, \ldots, J\}$. When $J = 1$, the model reduces to Example 1.

**Example 3** (Censored regression model)**.** Consider the model

$$Y_{it} = \max\{0, X_{it}^{\top}\beta_0 + U_{it}\},$$

where $\beta_0 \in \mathbb{R}^{d_x}$ is a vector of unknown coefficients. Here $\theta_0 = \beta_0$ and $\mathcal{Y} = [0, \infty)$.

**Example 4** (Multinomial choice model)**.** Suppose that $\mathcal{Y} = \{0, 1, \ldots, J\}$, and $X_{it}$ and $U_{it}$ consist of alternative-specific components:

$$X_{it} = (X_{0it}, X_{1it}, \ldots, X_{Jit}),\ U_{it} = (U_{0it}, U_{1it}, \ldots, U_{Jit}),$$

where for each $j$, $X_{jit} \in \mathbb{R}^k$ and $U_{jit} \in \mathbb{R}$. Consider the model

$$Y_{it} = \arg\max_j \, (X_{jit}^{\top}\beta_0 + U_{jit}),$$

where $\beta_0 \in \mathbb{R}^k$ is a vector of unknown coefficients. Here $\theta_0 = \beta_0$. Note that the normalization $\tilde{X}_{jit} = X_{jit} - X_{0it}, \tilde{U}_{jit} = U_{jit} - U_{0it}\ \forall j$ does not change outcomes. When $J = 1$, the model also reduces to Example 1.

**Assumption 1** (Time Homogeneity)**.** *$U_{it} \stackrel{d}{=} U_{i1} | X_i$ for all $t$.*

Assumption 1 requires that the conditional distribution of $U_{it}$ given $X_i$ does not depend on $t$. It is termed *time homogeneity* in Chernozhukov et al. (2013) and has been commonly imposed for semiparametric or nonparametric identification of nonlinear panel data models since its introduction by Manski (1987). A sufficient condition is that $U_{it}$ has an error component structure: $U_{it} = A_i + V_{it}$, where $V_{it} \stackrel{d}{=} V_{i1} | X_i, A_i$ for all $t$, and $A_i$ is a time-invariant individual effect. It is worth noting that Assumption 1 excludes lagged $Y_{it}$ from $X_{it}$ and focuses on static models. On the other hand, Assumption 1 allows $U_{it}$ to be correlated with $X_i$ and dependent over time. Moreover, it places no parametric distributional restriction on $U_{it}$.

**Assumption 2.** *$\theta_0$ is known or point-identified.*

Assumption 2 is satisfied for a broad class of structural functions $g$ under Assumption 1 and rich support conditions for $U_{it}$ and $X_i$. In particular, it holds for all the examples mentioned above. For Example 1, Manski (1987) showed the identification of $\beta_0$ up to scale. For Example 2, Botosaru et al. (2023) showed the identification of $\beta_0$ and $\gamma_0$ up to location and scale normalization by converting the model into a collection of binary choice models via binarization and invoking Manski (1987). For Example 3, Honoré and Kyriazidou (2000) showed the identification of $\beta_0$. For Example 4, point identification of $\theta_0$ up to scale is established in Shi et al. (2018) and Khan et al. (2021). Shi et al. (2018) exploited the cyclic monotonicity property of the choice probability vector. Khan et al. (2021) utilized the subsample of observations in which covariates for all alternatives but one are fixed over time to construct a localized rank-based objective function analogous to Manski (1987). Notably, a common structure is exploited by the identification argument of $\theta_0$ across these examples: $Y_{it}$ depends on $X_{it}$ and $U_{it}$ through latent indices $X_{it}^\top\beta_0 + U_{it}$ or $\{X_{jit}^\top\beta_0 + U_{jit}\}_{j=0}^J$. This structure will also be useful for the tractable implementation of sharp identified sets of counterfactuals in Section 4. However, it is not used in deriving sharp identified sets of counterfactuals in Section 3. In other words, results in Section 3 apply to more general settings that do not require this structure.

**Counterfactual Predictions** Fixing a counterfactual value $\underline{x}$ for $X_{it}$, the object of interest is the distribution of the counterfactual outcome $Y_{it}(\underline{x}) = g(\underline{x}, U_{it}; \theta_0)$. This can be understood as the result of an intervention that exogenously sets the value of $X_{it}$ to $\underline{x}$, without altering the structural function $g(\cdot;\theta_0)$ or the distribution of $U_{it}$. Summary measures of the distribution of $Y_{it}(\underline{x})$ can be formed in the spirit of the *average structural function* introduced in Blundell and Powell (2003, 2004). In Examples 2 and 3, one may consider the counterfactual survival probability $\Pr(Y_{it}(\underline{x}) \geq y)$ for $y \in \mathcal{Y} \setminus \inf \mathcal{Y}$. In Example 4, one may consider the counterfactual choice probability $\Pr(Y_{it}(\underline{x}) = y)$ for $y \in \mathcal{Y}$. These counterfactual probabilities are important parameters *per se* in evaluating the impact of counterfactual interventions. Moreover, they can serve as building blocks for various welfare measures. For example, Bhattacharya (2015, 2018) showed that in binary and multinomial choice models, the distribution of compensating and equivalent variation under a range of economic changes can be expressed as closed-form functionals of choice probabilities.

**Remark 1.** *The counterfactual evaluation point* $\underline{x}$ *can depend on* $X_i$*. For example,* $\underline{x}$ *can be the time average of* $X_i$ *shifted by a small amount. This allows for counterfactuals that fix the value of certain components of* $X_{it}$ *while leaving others at their realized values. However, I will omit this dependence for notational simplicity.*

**Remark 2.** *It may be interesting to consider counterfactuals that allow for endogenous responses to* $X_{it}$*, such as the imposition of a sales tax in supply-demand analysis. However, this requires a full structural model for the joint behavior of* $X_{it}$ *and* $U_{it}$ *and is beyond the scope of this paper.*

## 3 Identification

**Notation** For a generic random vector $W$, let $\mathcal{F}_{W|X} = \{F_{W|X=x} : x \in \text{Supp}(X)\}$ denote the collection of conditional distributions of $W$ given $X$, where for all $\mathcal{S} \subseteq \text{Supp}(W|X=x)$, $F_{W|X=x} = \Pr(W \in \mathcal{S}|X=x)$.

Define the $U$-level set as

$$\mathcal{U}(y_t, x_t; \theta) = \{u_t : y_t = g(x_t, u_t; \theta)\},$$

so that

$$u_t \in \mathcal{U}(y_t, x_t; \theta) \iff y_t = g(x_t, u_t; \theta).$$

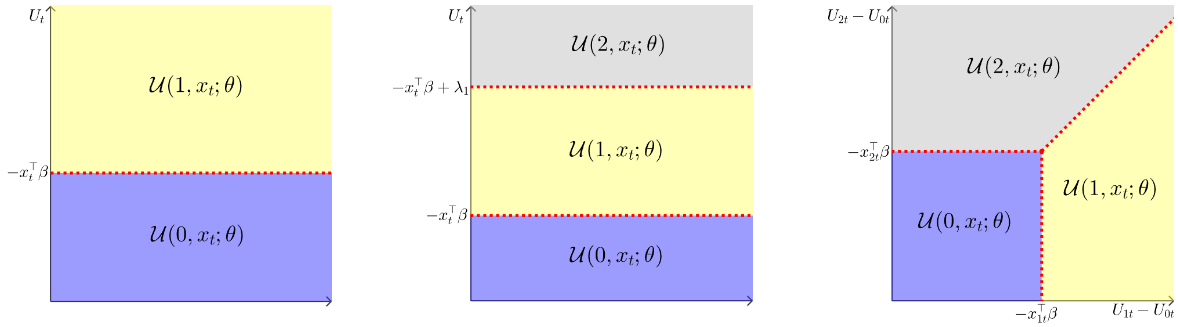
In words, $\mathcal{U}(y_t, x_t; \theta)$ denotes the set of values of $U_t$ that solves $Y_t = g(X_t, U_t; \theta)$ with structural function $g(\cdot; \theta)$ when $Y_t = y_t$ and $X_t = x_t$.[1] Figure 1 contains stylized depictions of $U$-level sets in Examples 1, 2, and 4 with $J = 2$. For any closed subset $\mathcal{T}$ of $\mathcal{Y}$, let $\mathcal{U}(\mathcal{T}, x_t; \theta) = \bigcup_{y_t \in \mathcal{T}} \mathcal{U}(y_t, x_t; \theta)$ so that $u_t \in \mathcal{U}(\mathcal{T}, x_t; \theta) \iff g(x_t, u_t; \theta) \in \mathcal{T}$.

Using $U$-level sets, the distribution of the counterfactual outcome $Y_t(\underline{x})$ can be characterized as

$$F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta_0)) \ \textit{a.e.}\ x \in \text{Supp}(X),\ \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}),$$

where $\mathsf{F}(\mathcal{Y})$ denotes the collection of all closed subsets of $\mathcal{Y}$. Therefore, to identify the distribution of $Y_t(\underline{x})$, it is necessary to identify $\theta_0$ and the distribution of $U_t|X=x$ over $\mathcal{U}(\mathcal{T}, \underline{x}; \theta_0)$ for each $\mathcal{T} \in \mathsf{F}(\mathcal{Y})$. The former, as discussed in Section 2, has been studied in the literature for a broad

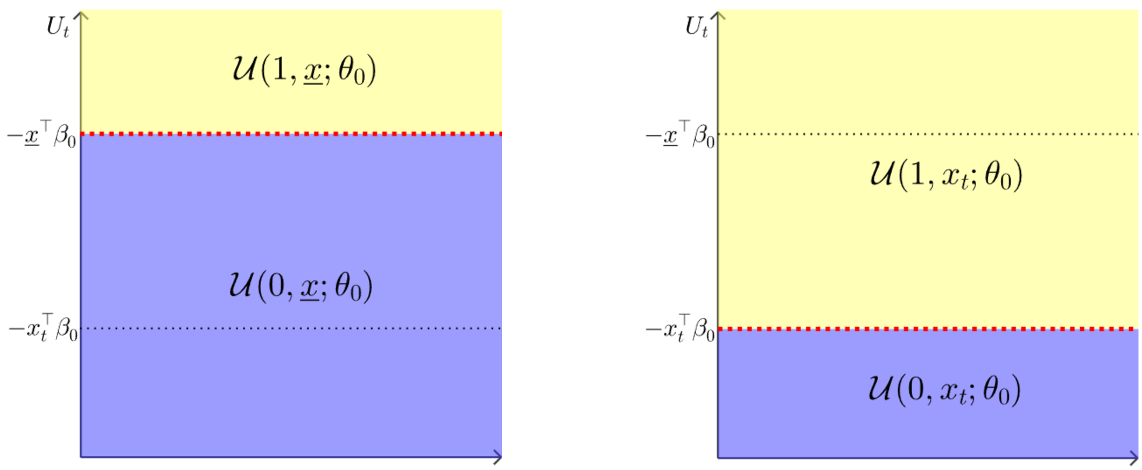
[1] To be clear, $\mathcal{U}(y_t, x_t; \theta)$ is merely the pre-image of $g(x_t, \cdot; \theta)$. I refer to it as the $U$-level set for simplicity, though it may be called by different names in other papers, such as the "disturbance region" in Pakes and Porter (2024).




(a) Binary Choice Model (b) Ordered Choice Model (c) Multinomial Choice Model

Figure 1: Stylized Depictions of $U$-Level Sets

class of nonlinear panel data models. The latter is a new element that emerges in the analysis of counterfactuals. When the outcome distribution exhibits mass points, such as in discrete or mixed distributions, point identification of both elements is impossible. I give a heuristic explanation for Example 1 using Figure 2.




(a) $U$-Level Sets under the Counterfactual (b) $U$-Level Sets for the Observed Data

Figure 2: Discrepancy of $U$-Level Sets: Binary Choice Model

As shown in Figure 2, for each $x \in \text{Supp}(X)$, the goal is to learn how $F_{U_t|X=x}$ allocates probability across $\mathcal{U}(1, \underline{x}; \theta_0)$ and $\mathcal{U}(0, \underline{x}; \theta_0)$. However, what is observed, $\Pr(Y_t = 1|X = x) = F_{U_t|X=x}(\mathcal{U}(1, x_t; \theta_0))$, only contains information about how probability is allocated across $\mathcal{U}(1, x_t; \theta_0)$ and $\mathcal{U}(0, x_t; \theta_0)$, which differ from $\mathcal{U}(1, \underline{x}; \theta_0)$ and $\mathcal{U}(0, \underline{x}; \theta_0)$ unless $\underline{x} = x_t$. Assumption 1 enables learning from $\Pr(Y_{t'} = 1|X = x)$ for $t' \neq t$ as well, but they may still lead to different $U$-level sets than desired. This discrepancy occurs for almost every $x \in \text{Supp}(X)$ if $X_t$ contains at least one continuous component, which is typically required for the point identification of $\theta_0$. As a result, the distribution of $U_t$ across $\mathcal{U}(1, \underline{x}; \theta_0)$ and $\mathcal{U}(0, \underline{x}; \theta_0)$ cannot be uniquely determined.

Given the impossibility of point identification, I provide the sharp identified set of the distribution of $Y_t(\underline{x})$ in Theorem 1. The proof is in Appendix A. The sharp identified set relies on the standard definition of *observational equivalence*, that is, it collects all the distributions of $Y_t(\underline{x})$ that can be reproduced by a distribution of $U_t$ consistent with the observed data. A key simplification afforded by Assumption 1 is that, although one observes joint distributions $\mathcal{F}_{Y|X}$, the distribution of $U_t$ is only required to match the marginals $\{\mathcal{F}_{Y_{t'}|X}\}_{t'=1}^{T}$, and one can combine these restrictions by taking intersection across $t'$. In this sense, a long panel plays an analogous role to that of an instrument with rich variation.

**Theorem 1.** *Suppose that Assumptions 1 and 2 hold. Then, the sharp identified set for $\mathcal{F}_{Y_t(\underline{x})|X}$, denoted by $\mathsf{F}^*_{Y_t(\underline{x})|X}$, is given by*

$$\begin{aligned}\mathsf{F}^*_{Y_t(\underline{x})|X} = \{&\mathcal{F}_{Y_t(\underline{x})|X} : \exists \mathcal{F}_{U_t|X} \in \mathsf{F}^*_{U_t|X} \\ &\text{s.t. } \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta_0)) \text{ a.e. } x \in \operatorname{Supp}(X)\},\end{aligned} \tag{1}$$

*where $\mathsf{F}^*_{U_t|X}$ collects the distributions of $U_t$ consistent with the observed data in the sense that*

$$\mathsf{F}^*_{U_t|X} = \bigcap_{t'=1}^{T} \{\mathcal{F}_{U_t|X} : \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_{t'}|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, x_{t'}; \theta_0)) \text{ a.e. } x \in \operatorname{Supp}(X)\}.$$

**Remark 3.** *Point identification of $\theta_0$ (Assumption 2) is imposed to fix ideas and is stronger than necessary. The identified set defined in (1) is sharp for a given value of $\theta_0$. When point identification of $\theta_0$ fails, one can still take the union of (1) over the sharp identified set for $\theta_0$ to obtain the sharp identified set for $\mathcal{F}_{Y_t(\underline{x})|X}$.*

# 4 Implementation

By Theorem 1, the most straightforward way to implement $\mathsf{F}^*_{Y_t(\underline{x})|X}$ is to search over the space of distributions supported on

$$\mathbb{U}(x) = \Big\{\mathcal{U}(y, \underline{x}; \theta_0) \cap \Big(\bigcap_{t'=1}^{T} \mathcal{U}(y_{t'}, x_{t'}; \theta_0)\Big) : (y, y_1, \ldots, y_T) \in \mathcal{Y}^{T+1}\Big\}$$

for each $x \in \text{Supp}(X)$. With discrete outcomes, $\mathbb{U}(x)$ is a finite partition of the space of $U_t$, and any point within each set in $\mathbb{U}(x)$ produces the same outcome under $\underline{x}, x_1, \ldots, x_T$. This extends the concept of the *minimal relevant partition* of Tebaldi et al. (2023) to general discrete choice models.[2] Nonetheless, depending on $T$, the cardinality of $\mathcal{Y}$, and the structural function $g$, the cardinality of $\mathbb{U}(x)$ can be large, making the search computationally demanding. In this section, I provide tractable characterizations of $\mathsf{F}^*_{Y_t(\underline{x})|X}$ that avoid directly searching over the distributions of $U_t$ by exploiting the separable index restriction on $g$, with a focus on Examples 1-4. I start with a heuristic illustration in Example 1.

As shown in Figure 2, $U$-level sets are half intervals: $\mathcal{U}(1, x_t; \theta_0) = [-x_t^\top \beta_0, \infty)$. Hence, when the value of explanatory variables is changed from observed to counterfactual ones, there is a set inclusion relationship between the corresponding $U$-level sets, which can be translated into a comparison between the distributions of observed and counterfactual outcomes:

$$\begin{aligned}
\underline{x}^\top \beta_0 \leq x_t^\top \beta_0 \quad &\Longleftrightarrow \quad \mathcal{U}(1, \underline{x}; \theta_0) \subseteq \mathcal{U}(1, x_t; \theta_0) \iff F_{Y_t(\underline{x})|X=x}(\{1\}) \leq F_{Y_t|X=x}(\{1\}), \\
\underline{x}^\top \beta_0 \geq x_t^\top \beta_0 \quad &\Longleftrightarrow \quad \mathcal{U}(1, \underline{x}; \theta_0) \supseteq \mathcal{U}(1, x_t; \theta_0) \iff F_{Y_t(\underline{x})|X=x}(\{1\}) \geq F_{Y_t|X=x}(\{1\}).
\end{aligned}$$

In this way, I generate identifying restrictions on $\mathcal{F}_{Y_t(\underline{x})|X}$ directly from $\mathcal{F}_{Y_t|X}$. Under Assumption 1, I can repeat this procedure using observed data from any period. The resulting identifying restrictions turn out to be sharp.

Beyond binary choice models, set inclusion relationships of $U$-level sets generally take the form

$$\mathcal{U}(\mathcal{T}, \underline{x}; \theta_0) \subseteq \mathcal{U}(\mathcal{T}', x_t; \theta_0)$$

for some $\mathcal{T}, \mathcal{T}' \in \mathsf{F}(\mathcal{Y})$, implying that

$$F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \leq F_{Y_t|X=x}(\mathcal{T}').$$

As previewed at the end of Section 2, a common structure in Examples 1-4 makes it easier to

[2] I present the formal definition of the minimal relevant partition here for completeness. Let $\mathcal{X}_r$ denote a finite set of relevant values of explanatory variables, which can contain both observed and counterfactual values. The minimal relevant partition is a collection $\mathbb{U}$ of sets $\mathcal{U} \in \mathbb{R}^{d_u}$ for which the following property holds for almost every $u, u' \in \mathbb{R}^{d_u}$ (with respect to Lebesgue measure):

$$u, u' \in \mathcal{U} \text{ for some } \mathcal{U} \in \mathbb{U} \iff g(\tilde{x}, u; \theta_0) = g(\tilde{x}, u'; \theta_0) \text{ for all } \tilde{x} \in \mathcal{X}_r.$$

Then, $\mathbb{U}(x)$ is a minimal relevant partition by letting $\mathcal{X}_r = \{\underline{x}, x_1, \ldots, x_T\}$.

determine these set inclusion relationships. More formally, Examples 1-4 satisfy an *index separability* condition in the sense that by partitioning $\theta = (\beta, \gamma)$,

$$(\mathcal{T}, \mathcal{T}') \in \mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma) \;\Rightarrow\; \mathcal{U}(\mathcal{T}, \underline{x}; \theta) \subseteq \; \mathcal{U}(\mathcal{T}', x_t; \theta) \tag{2}$$

for some collection $\mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma)$ of pairs of subsets of $\mathcal{Y}$.[3] In words, (2) means that the set inclusion relationship between $\mathcal{U}(\mathcal{T}, \underline{x}; \theta)$ and $\mathcal{U}(\mathcal{T}', x_t; \theta)$ can be determined by examining the pair of indices $(\underline{x}^\top \beta, x_t^\top \beta)$. By carefully selecting $\mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma)$, the implied set inclusion relationships of $U$-level sets can be shown to exhaust all the information on the distribution of $Y_t(\underline{x})$.

Examples 1-3 are encompassed by the following *monotone transformation model.*

**Example 5** (Monotone Transformation Model)**.** Consider the model

$$Y_t = h(X_t^\top \beta_0 + U_t; \gamma_0),$$

where $\beta_0 \in \mathbb{R}^{d_x}$ is a vector of unknown coefficients, and $h$ is a transformation function that is weakly increasing, right-continuous, and known up to a finite-dimensional parameter $\gamma_0$. For Example 1, $h(v; \gamma) = 1\{v \geq 0\}$. For Example 2, $h(v; \gamma) = \sum_{j=0}^{J} 1\{v \geq \gamma^j\}$. For Example 3, $h(v; \gamma) = \max\{0, v\}$. Define the generalized inverse of $h$ as

$$h^-(y; \gamma) = \inf\{y^* \in \mathcal{Y} : h(y^*; \gamma) \geq y\}, \; y \in \mathcal{Y}.$$

Then, $U$-level sets satisfy

$$\mathcal{U}([y, \infty), x_t; \theta) = [-x_t^\top \beta + h^-(y; \gamma), \infty). \tag{3}$$

Also define

$$\begin{aligned}
\mathbb{Y}_u(\underline{x}^\top \beta, x_t^\top \beta; \gamma) &= \{([y, \infty), [y', \infty)) \cap \mathcal{Y}^2 : (y, y') \in \mathcal{Y}, -\underline{x}^\top \beta + h^-(y; \gamma) \geq -x_t^\top \beta + h^-(y'; \gamma)\}, \\
\mathbb{Y}_l(\underline{x}^\top \beta, x_t^\top \beta; \gamma) &= \{([y, \infty), [y', \infty)) \cap \mathcal{Y}^2 : (y, y') \in \mathcal{Y}, -\underline{x}^\top \beta + h^-(y; \gamma) \leq -x_t^\top \beta + h^-(y'; \gamma)\}.
\end{aligned}$$

[3] A more general form allowing for nonlinear indices replaces $\mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma)$ with $\mathbb{Y}(s(\underline{x}, \theta), s(x_t, \theta); \theta)$, where $s(\cdot; \theta)$ is a potentially vector-valued function known up to $\theta$. However, in this paper, I focus on linear indices that are the most commonly used in practice.

One can predict the following set inclusion relationships:

$$(\mathcal{T},\mathcal{T}') \in \mathbb{Y}_u(\underline{x}^\top\beta, x_t^\top\beta;\gamma) \quad\Longleftrightarrow\quad \mathcal{U}(\mathcal{T},\underline{x};\theta) \subseteq \mathcal{U}(\mathcal{T}',x_t;\theta),$$
$$(\mathcal{T},\mathcal{T}') \in \mathbb{Y}_l(\underline{x}^\top\beta, x_t^\top\beta;\gamma) \quad\Longleftrightarrow\quad \mathcal{U}(\mathcal{T},\underline{x};\theta) \supseteq \mathcal{U}(\mathcal{T}',x_t;\theta).$$

**Example 4** (continued)**.** Note that for any $\mathcal{T} \subsetneq \{0,1,\ldots,J\}$ such that $\mathcal{T} \neq \emptyset$,

$$\mathcal{U}(\mathcal{T},x_t;\theta) = \Big\{U_t : \max_{j\in\mathcal{T}} x_{jt}^\top\beta + U_{jt} \geq \max_{k\notin\mathcal{T}} x_{kt}^\top\beta + U_{kt}\Big\}.$$

Since $\gamma$ is not present in this example, I omit it and define

$$\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) = \Big\{(\mathcal{T},\mathcal{T}) : \mathcal{T} \subsetneq \{0,1,\ldots,J\}, \mathcal{T}\neq\emptyset, \min_{j\in\mathcal{T}}(x_{jt}-\underline{x}_j)^\top\beta \geq \max_{k\notin\mathcal{T}}(x_{kt}-\underline{x}_k)^\top\beta\Big\}. \quad (4)$$

Intuitively, for any $\mathcal{T}$ satisfying the restrictions in (4), moving from $\underline{x}$ to $x_t$ makes alternatives in $\mathcal{T}$ more likely to be chosen, regardless of the distribution of $U_t$. Hence, one can predict the following set inclusion relationships:

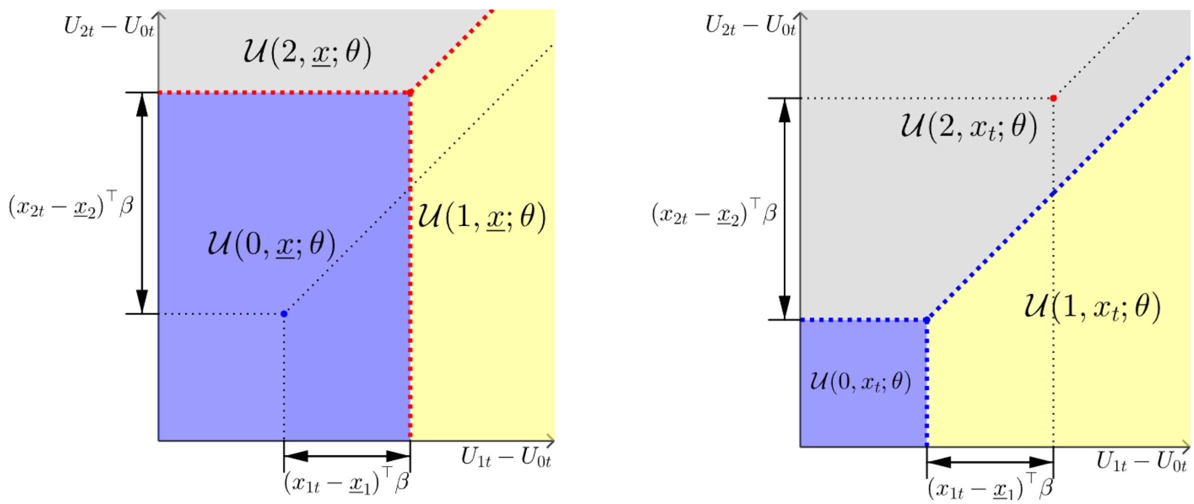
$$(\mathcal{T},\mathcal{T}') \in \mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) \;\Rightarrow\; \mathcal{U}(\mathcal{T},\underline{x};\theta) \subseteq \mathcal{U}(\mathcal{T}',x_t;\theta). \quad (5)$$

A proof of relation (5) is given in Appendix A. It is helpful to understand (5) graphically. Consider the case of $J = 2$ and suppose that $(x_{2t}-\underline{x}_2)^\top\beta > (x_{1t}-\underline{x}_1)^\top\beta > 0$. Then, $\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) = \{\{2\},\{2,1\}\}$. As shown in Figure 3, there are two set inclusion relationships:

$$\mathcal{U}(2,\underline{x};\theta) \;\subseteq\; \mathcal{U}(2,x_t;\theta),$$
$$\mathcal{U}(2,\underline{x};\theta)\cup\mathcal{U}(1,\underline{x};\theta) \;\subseteq\; \mathcal{U}(2,x_t;\theta)\cup\mathcal{U}(1,x_t;\theta).$$

In general, to construct $\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta)$, one can simply rank the $J+1$ index function differences $\{(x_{jt}-\underline{x}_j)^\top\beta\}_{j=0}^J$ and collect the $\mathcal{T}$'s that contain the top $j$ alternatives for $j = 1,\ldots J$.

With the set inclusion relationships of $U$-level sets discussed above, I am ready to present tractable characterizations of $\mathsf{F}^*_{Y_t(\underline{x})|X}$ for Examples 5 and 4 in Theorems 2 and 3, respectively. The proofs are in Appendix A.




(a) $U$-Level Sets under the Counterfactual (b) $U$-Level Sets for the Observed Data

Figure 3: Set Inclusion Relationships of $U$-Level Sets: Multinomial Choice Model

**Theorem 2.** *Suppose that Assumptions 1 and 2 hold. Let $g$ be specified as in Example 5. Then,*

$$\begin{aligned}\mathsf{F}^*_{Y_t(\underline{x})|X} = \bigcap_{t'=1}^{T} \Big\{ &\mathcal{F}_{Y_t(\underline{x})|X} : \forall(\mathcal{T},\mathcal{T}') \in \mathbb{Y}_u(\underline{x}^\top\beta_0, x_{t'}^\top\beta_0;\gamma_0), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \leq F_{Y_{t'}|X=x}(\mathcal{T}'),\\ &\forall(\mathcal{T},\mathcal{T}') \in \mathbb{Y}_l(\underline{x}^\top\beta_0, x_{t'}^\top\beta_0;\gamma_0), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \geq F_{Y_{t'}X=x}(\mathcal{T}') \text{ a.e. } x \in \mathrm{Supp}(X)\Big\}. \end{aligned} \tag{6}$$

By Theorem 2, the sharp bounds on the counterfactual survival probability $F_{Y_t(\underline{x})|X=x}([y,\infty))$ are given by

$$\bigcap_{t'=1}^{T}\Bigg[\sup_{\substack{y':-\underline{x}^\top\beta_0+h^-(y;\gamma_0)\\ \geq -x_{t'}^\top\beta_0+h^-(y';\gamma_0)}} F_{Y_{t'}|X=x}([y',\infty)), \inf_{\substack{y':-\underline{x}^\top\beta_0+h^-(y;\gamma_0)\\ \leq -x_{t'}^\top\beta_0+h^-(y';\gamma_0)}} F_{Y_{t'}|X=x}([y',\infty))\Bigg]$$

with the convention that $\sup\emptyset = 0$ and $\inf\emptyset = 1$. This result is similar to Theorem 2 of Botosaru and Muris (2024), where they allow the transformation function $h$ to vary over time. My framework can also accommodate time-varying $h$ as long as it is point-identified. The key difference is that I establish the sharpness of their bounds.

**Theorem 3.** *Suppose that Assumptions 1 and 2 hold. Let $g$ be specified as in Example 4. Then,*

$$\mathsf{F}^*_{Y_t(\underline{x})|X} = \bigcap_{t'=1}^{T}\{\mathcal{F}_{Y_t(\underline{x})|X} : \forall(\mathcal{T},\mathcal{T}') \in \mathbb{Y}(\underline{x}^\top\beta_0, x_{t'}^\top\beta_0), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \leq F_{Y_{t'}|X=x}(\mathcal{T}') \text{ a.e. } x \in \mathrm{Supp}(X)\}. \tag{7}$$

A collection of choice sets similar to (4) appears in Pakes and Porter (2024). They used the set inclusion relationship of $U$-level sets for the observed data between two time periods to derive identifying restrictions on the structural parameter $\theta_0$. They also showed that when $T = 2$, these identifying restrictions are sharp and yield point identification under the additional conditions given in Shi et al. (2018). My results further open up the possibility of counterfactual analysis built upon the knowledge of $\theta_0$.

## 5 Estimation and Inference

In this section, I focus on discrete outcomes. Let

$$\tau_0(x) = \{F_{Y_t|X=x}(\{y\}) : y \in \mathcal{Y}, t \in \{1, \ldots, T\}\}$$

denote the vector of observed conditional choice probabilities. I consider estimation and inference of aggregated intersection bounds that can be written as

$$[\Psi_l(\theta_0), \Psi_u(\theta_0)] = \Big[E\Big[\max_{\lambda \in \Lambda_l(X;\theta_0)} \lambda^\top \tau_0(X)\Big], E\Big[\min_{\lambda \in \Lambda_u(X;\theta_0)} \lambda^\top \tau_0(X)\Big]\Big], \tag{8}$$

where $\Lambda_l(x;\theta)$ and $\Lambda_u(x;\theta)$ are known finite sets, and expectations are taken over $X$. The reason is that the bounds on summary measures of the counterfactual outcome distribution can be expressed as in (8). I demonstrate this point with examples. For $\mathcal{T} \subseteq \mathcal{Y}$, let $e_{\mathcal{T}} \in \{0,1\}^{|\mathcal{Y}|}$ be a vector whose $y$th component is 1 if $y \in \mathcal{T}$. For $t \in \{1, \ldots, T\}$, let $e_t$ be a unit vector with 1 in its $t$th place.

**Example 2** (continued)**.** Fixing a counterfactual value $\underline{x}$ for $X_t$, the sharp bounds on counterfactual survival probabilities $\Pr(Y_t(\underline{x}) \geq j)$ take the form of (8). To see this, note that by Theorem 2, the bounds are given by $[E[\max_t \psi_t^l(X;\theta_0)], E[\min_t \psi_t^u(X;\theta_0)]]$, where

$$\begin{aligned}
\psi_t^l(x;\theta) &= F_{Y_t|X=x}(\{k : k \geq \min\{y \in \mathcal{Y} : -\underline{x}^\top \beta + h^-(j;\gamma) \leq -x_t^\top \beta + h^-(y;\gamma)\}\}), \\
\psi_t^u(x;\theta) &= F_{Y_t|X=x}(\{k : k \geq \max\{y \in \mathcal{Y} : -\underline{x}^\top \beta + h^-(j;\gamma) \geq -x_t^\top \beta + h^-(y;\gamma)\}\}),
\end{aligned}$$

with the convention that $\min \emptyset = \infty$. Define

$$\begin{aligned}
\mathcal{T}_t^l(x;\theta) &= \{k : k \geq \min\{y \in \mathcal{Y} : -\underline{x}^\top \beta + h^-(j;\gamma) \leq -x_t^\top \beta + h^-(y;\gamma)\}\}, \\
\mathcal{T}_t^u(x;\theta) &= \{k : k \geq \max\{y \in \mathcal{Y} : -\underline{x}^\top \beta + h^-(j;\gamma) \geq -x_t^\top \beta + h^-(y;\gamma)\}\}.
\end{aligned}$$

Then, $\psi_t^l(x;\theta)$ and $\psi_t^u(x;\theta)$ can be written as linear functions of $\tau_0(x)$:

$$\psi_t^l(x;\theta) = (e_t \otimes e_{\mathcal{T}_t^l(x;\theta)})^\top \tau_0(x), \quad \psi_t^u(x;\theta) = (e_t \otimes e_{\mathcal{T}_t^u(x;\theta)})^\top \tau_0(x).$$

Now define

$$\Lambda_l(x;\theta) = \{e_t \otimes e_{\mathcal{T}_t^l(x;\theta)} : t \in \{1,\ldots,T\}\}, \quad \Lambda_u(x;\theta) = \{e_t \otimes e_{\mathcal{T}_t^u(x;\theta)} : t \in \{1,\ldots,T\}\}.$$

Then,

$$E[\max_t \psi_t^l(X;\theta_0)] = E\Big[\max_{\lambda \in \Lambda_l(x;\theta_0)} -\lambda^\top \tau_0(X)\Big], \quad E[\min_t \psi_t^u(X;\theta_0)]] = E\Big[\min_{\lambda \in \Lambda_u(x;\theta_0)} \lambda^\top \tau_0(X)\Big].$$

**Example 4** (continued)**.** Fixing a counterfactual value $\underline{x}$ for $X_t$, the sharp bounds on counterfactual choice probabilities $\Pr(Y_t(\underline{x}) = j)$ take the form of (8). To see this, note that by Theorem 3, the bounds are given by $[E[\max_t \psi_t^l(X;\theta_0)], E[\min_t \psi_t^u(X;\theta_0)]]$, where $\psi_t^l(x;\theta)/\psi_t^u(x;\theta)$ is the solution to the linear program

$$\begin{aligned} &\max / \min_{\vec{q} \in \Delta^{J+1}} \ q_j \\ &\text{s.t.} \sum_{j \in \mathcal{T}} q_j \leq F_{Y_{t'}|X=x}(\mathcal{T}') \ \forall (\mathcal{T}, \mathcal{T}') \in \mathbb{Y}(\underline{x}^\top \beta, x_{t'}^\top \beta), \forall t' \in \{1,\ldots,T\}, \end{aligned}$$

where $\Delta^{J+1}$ denotes the probability simplex in $\mathbb{R}^{J+1}$. Some algebra reveals that $\psi_t^l(x;\theta)$ and $\psi_t^u(x;\theta)$ have closed forms:

$$\begin{aligned} \psi_t^l(x;\theta) &= \begin{cases} F_{Y_t|X=x}(\{j\}) & \text{if } (x_{jt} - \underline{x}_j)^\top \beta \leq (x_{kt} - \underline{x}_k)^\top \beta, \ \forall k \\ 0 & \text{otherwise} \end{cases}, \\ \psi_t^u(x;\theta) &= \begin{cases} F_{Y_t|X=x}(\{j\}) & \text{if } (x_{jt} - \underline{x}_j)^\top \beta \geq (x_{kt} - \underline{x}_k)^\top \beta, \ \forall k \\ F_{Y_t|X=x}(\{j\} \cup \{k : (x_{kt} - \underline{x}_k)^\top \beta > (x_{jt} - \underline{x}_j)^\top \beta\}) & \text{otherwise} \end{cases}. \end{aligned}$$

Define

$$\mathcal{T}_t^l(x;\theta) = \begin{cases} \{j\} & \text{if } (x_{jt} - \underline{x}_j)^\top \beta \leq (x_{kt} - \underline{x}_k)^\top \beta, \ \forall k \\ \emptyset & \text{otherwise} \end{cases},$$

$$\mathcal{T}_t^u(x;\theta) \;=\; \begin{cases} \{j\} & \text{if } (x_{jt}-\underline{x}_j)^\top\beta \geq (x_{kt}-\underline{x}_k)^\top\beta,\ \forall k \\ \{j\}\cup\{k:(x_{kt}-\underline{x}_k)^\top\beta > (x_{jt}-\underline{x}_j)^\top\beta\} & \text{otherwise} \end{cases}.$$

It is again evident that $\psi_t^l(x;\theta)$ and $\psi_t^u(x;\theta)$ are linear functions of $\tau_0(x)$:

$$\psi_t^l(x;\theta) = (e_t \otimes e_{\mathcal{T}_t^l(x;\theta)})^\top \tau_0(x), \quad \psi_t^u(x;\theta) = (e_t \otimes e_{\mathcal{T}_t^u(x;\theta)})^\top \tau_0(x).$$

Then, the argument used in the previous example applies.

To construct estimators of $\Psi_l(\theta_0)$ and $\Psi_u(\theta_0)$, I use cross-fitting to estimate $\tau_0$.

**Definition 1** (Cross-fitting)**.** Divide the cross-sectional units into $K$ evenly-sized folds. For each $k = 1, \ldots, K$, use the other $K-1$ folds to estimate $\tau_0$; denote the resulting estimates by $\hat{\tau}^{(-k)}$. For each $i = 1, \ldots, N$, take $\hat{\tau}(X_i) = \hat{\tau}^{(-k_i)}(X_i)$, where $k_i$ denotes the fold containing the $i$th observation.

Let $\|\cdot\|$ denote the Euclidean norm. I impose the following assumptions.

**Assumption 3.** *For all $\theta$, $\max_{\lambda\in\Lambda_l(x;\theta)\cup\Lambda_u(x;\theta)} \|\lambda\| \leq M$ for some $M > 0$ a.e. $x \in \mathrm{Supp}(X)$.*

**Assumption 4.** *For all $\theta$ and $\tau$, $\arg\max_{\lambda\in\Lambda_l(x;\theta)} \lambda^\top\tau(x)$ and $\arg\min_{\lambda\in\Lambda_u(x;\theta)} \lambda^\top\tau(x)$ are singletons a.e. $x \in \mathrm{Supp}(X)$.*

**Assumption 5.** *The distribution of $\tau_0(X)$ is absolutely continuous with density bounded above.*

**Assumption 6.** *$\|\hat{\tau} - \tau_0\|_\infty = o_p(N^{-1/4})$, where $\|\tau\|_\infty = \sup_x \|\tau(x)\|$.*

Assumption 3 imposes boundedness on the objective function of the optimization problems and is satisfied in Examples 2 and 4. Assumption 4 requires the solution of the optimization problems to be unique. Assumption 5 is a sufficient condition for the margin condition (Lemma 1) that controls the concentration of the objective function in the neighborhood of the optimum. In other words, it ensures the optimum is separated from non-optimal values with high probability. The uniqueness of the optimal solution and the margin condition are also imposed in Semenova (2024) to derive inference for a general class of aggregated intersection bounds. I retain Assumption 5 because it is low-level and compatible with the sufficient conditions for Assumption 2. Assumption 6 requires the estimation error of $\hat{\tau}$ to vanish fast enough. The $o_p(N^{-1/4})$ rate is a classic assumption in the semiparametric estimation literature. One may use the series logit estimator in Hirano, Imbens, and Ridder (2003). Let

$$I(Y) = \{1\{Y_t = y\} : y \in \mathcal{Y}, t \in \{1, \ldots, T\}\}$$

be a vector of binary indicators that is conformable with $\tau_0(x)$. Define

$$\lambda_l^*(x;\theta,\tau) = \underset{\lambda \in \Lambda_l(x;\theta)}{\arg\max}\, \lambda^\top \tau(x), \quad \lambda_u^*(x;\theta,\tau) = \underset{\lambda \in \Lambda_u(x;\theta)}{\arg\min}\, \lambda^\top \tau(x).$$

Given the first-step cross-fitted estimator $\hat{\tau}$ of $\tau_0$, define

$$\begin{aligned} \hat{\Psi}_l(\theta) &= \frac{1}{N}\sum_{i=1}^{n} \sum_{\lambda \in \Lambda_l(X_i;\theta)} 1\{\lambda_l^*(X_i;\theta,\hat{\tau}) = \lambda\}\lambda^\top I(Y_i), \\ \hat{\Psi}_u(\theta) &= \frac{1}{N}\sum_{i=1}^{n} \sum_{\lambda \in \Lambda_u(X_i;\theta)} 1\{\lambda_u^*(X_i;\theta,\hat{\tau}) = \lambda\}\lambda^\top I(Y_i). \end{aligned}$$

**Theorem 4.** *Suppose that Assumptions 3-6 hold. Then, for a given $\theta$,*

$$\begin{aligned} \sqrt{N}(\hat{\Psi}_l(\theta) - \Psi_l(\theta)) &\xrightarrow{d} N(0, V_l(\theta)), \\ \sqrt{N}(\hat{\Psi}_u(\theta) - \Psi_u(\theta)) &\xrightarrow{d} N(0, V_u(\theta)), \end{aligned}$$

*where*

$$\begin{aligned} V_l(\theta) &= E\Big[\sum_{\lambda \in \Lambda_l(X;\theta)} 1\{\lambda_l^*(X;\theta,\tau_0) = \lambda\}(\lambda^\top I(Y))^2\Big] - \Psi_l^2(\theta), \\ V_u(\theta) &= E\Big[\sum_{\lambda \in \Lambda_u(X;\theta)} 1\{\lambda_u^*(X;\theta,\tau_0) = \lambda\}(\lambda^\top I(Y))^2\Big] - \Psi_u^2(\theta). \end{aligned}$$

In view of Theorem 4, a natural idea is to plug in a first-step estimate $\hat{\theta}$ of $\theta_0$ to obtain the final estimators $\hat{\Psi}_l(\hat{\theta})$ and $\hat{\Psi}_u(\hat{\theta})$. However, the asymptotic distribution of such plug-in estimators is complicated by the estimation error of $\hat{\theta}$. I give a heuristic explanation for $\hat{\Psi}_l(\hat{\theta})$ in Example 1. One can decompose

$$\hat{\Psi}_l(\hat{\theta}) - \Psi_l(\theta_0) = \hat{\Psi}_l(\hat{\theta}) - \Psi_l(\hat{\theta}) + \Psi_l(\hat{\theta}) - \Psi_l(\theta_0).$$

By Theorem 4, $\hat{\Psi}_l(\hat{\theta}) - \Psi_l(\hat{\theta}) = O(N^{-1/2})$. Note that $\theta$ enters $\Psi_l(\theta)$ only through $\Lambda_l$ so that

$$|\Psi_l(\hat{\theta}) - \Psi_l(\theta_0)| = O(\Pr(\Lambda_l(X;\hat{\theta}) \neq \Lambda_l(X;\theta_0))).$$

For $\theta \neq \theta_0$, $\Lambda_l(x;\theta) \neq \Lambda_l(x;\theta_0)$ if for some $t$, $\operatorname{sgn}((x_t - \underline{x})^\top \beta) \neq \operatorname{sgn}((x_t - \underline{x})^\top \beta_0)$, which occurs with probability of order $O(\|\theta - \theta_0\|)$. Therefore, $\Psi_l(\hat{\theta}) - \Psi_l(\theta_0)$ becomes dominating in the expansion of $\hat{\Psi}_l(\hat{\theta})$ if $\hat{\theta}$ converges at a slower rate than $N^{-1/2}$, as is the case with the maximum estimator proposed by Manski (1987) and its smoothed version.

To utilize the asymptotic normality result in Theorem 4, I consider Bonferroni-type confidence intervals. To this end, define

$$\begin{aligned}\hat{V}_l(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{\lambda\in\Lambda_l(X_i;\theta)} 1\{\lambda_l^*(X_i;\theta,\hat{\tau})=\lambda\}(\lambda^\top I(Y_i))^2 - \hat{\Psi}_l^2(\theta),\\ \hat{V}_u(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{\lambda\in\Lambda_u(X_i;\theta)} 1\{\lambda_u^*(X_i;\theta,\hat{\tau})=\lambda\}(\lambda^\top I(Y_i))^2 - \hat{\Psi}_u^2(\theta),\end{aligned}$$

which are consistent estimators of $V_l(\theta)$ and $V_u(\theta)$ for a given $\theta$ under Assumption 6. Also, suppose that one can construct a $(1-\delta)$-confidence region for $\theta_0$:

$$\lim_{N\to\infty}\Pr(\theta_0\in \mathrm{CR}_N(\delta)) = 1-\delta. \tag{9}$$

Construction of $\mathrm{CR}_N(\alpha)$ is possible using existing estimators of $\theta_0$. A brief review is provided below. For $0\le\delta<\alpha$, the Bonferroni confidence interval for $[\Psi_l(\theta_0),\Psi_u(\theta_0)]$ is given by

$$\mathrm{CI}_N(\alpha,\delta) = \Big[\inf_{\theta\in\mathrm{CR}_N(\delta)}\hat{\Psi}_l(\theta) - z_{1-(\alpha-\delta)/2}\sqrt{\hat{V}_l(\theta)/N},\ \sup_{\theta\in\mathrm{CR}_N(\delta)}\hat{\Psi}_u(\theta) + z_{1-(\alpha-\delta)/2}\sqrt{\hat{V}_u(\theta)/N}\Big].$$

**Proposition 1.** *Suppose that Assumptions 3-6 and (9) hold. Then, for any* $0\le\delta<\alpha$,

$$\lim_{N\to\infty}\Pr([\Psi_l(\theta_0),\Psi_u(\theta_0)]\subseteq \mathrm{CI}_N(\alpha,\delta)) = 1-\alpha.$$

In Appendix B, I conduct a Monte Carlo experiment to evaluate the performance of the confidence interval in Proposition 1.

**Remark 4.** *The confidence interval in Proposition 1 is for the sharp identified set* $[\Psi_l(\theta_0),\Psi_u(\theta_0)]$ *of the counterfactual probability, not the counterfactual probability itself. If the latter is of interest, one may adapt the methods of Imbens and Manski (2004) and Stoye (2009) to construct confidence intervals that are less conservative yet uniformly valid, but this is beyond the scope of this paper.*

**Remark 5.** *The confidence interval in Proposition 1 is two-sided. If one is only interested in the upper or lower bound on the counterfactual probability, it is straightforward to construct a one-sided confidence interval by using* $z_{1-(\alpha-\delta)}$ *instead of* $z_{1-(\alpha-\delta)/2}$ *and setting the other side to* $-\infty$ *or* $\infty$.

The literature on semiparametric inference for $\theta_0$ has not yet converged on a single procedure. For panel data binary choice models, the asymptotic distribution of the maximum score estimator is that of the maximizer of a Gaussian process, which is hard to use for inference. One solution is to switch to the smoothed maximum score estimator proposed by Charlier, Melenberg, and van Soest (1995), but this requires selecting an additional kernel function and tuning parameters. An alternative is to use bootstrap-based methods. Abrevaya and Huang (2005) have shown that the classic bootstrap is inconsistent for the maximum score estimators. Valid inference may be conducted using subsampling (Delgado, Rodríguez-Poo, and Wolf, 2001), $m$-out-of-$n$ bootstrap (Lee and Pun, 2006), the numerical bootstrap (Hong and Li, 2020), and a model-based bootstrap procedure that analytically modifies the criterion function (Cattaneo, Jansson, and Nagasawa, 2020). For panel data multinomial choice models, Khan et al. (2021) proposed a localized maximum score estimator, whose asymptotic distribution is also that of the maximizer of a Gaussian process. Khan et al. (2021) conjectured that both a smoothed maximum score approach and bootstrap-based procedures may be used for inference.

# 6 Numerical Experiments

In this section, I investigate how identifying power varies with the number of time periods and the cardinality of outcome support through numerical experiments.

For Example 2, I consider the following data generating process:

$$Y_t = \sum_{j=0}^{J} 1\{\beta_0^{(1)} X_t^{(1)} + \beta_0^{(2)} X_t^{(2)} + U_t \geq \gamma_0^j\},\ t = 1, \dots, T,$$

where $X_t^{(1)} \sim N(0, 0.5)$ and $U_t = A + V_t$ with $V_t \sim N(0, 0.5)$. I define two equally sized latent populations of cross-sectional units. In the first population, $X_t^{(2)} \sim \text{Bernoulli}(0.5)$ and $A = 1 + (0.5 + T(\bar{X}^{(1)})^2) \cdot Z$, while in the second population, $X_t^{(2)} = 0$ and $A = (0.5 + T(\bar{X}^{(1)})^2) \cdot Z$, where $\bar{X}^{(1)} = \frac{1}{T} \sum_{t=1}^{T} X_t^{(1)}$ and $Z \sim N(0, 0.5)$. In summary, $A$ is heteroskedastic, with its variance depending on $X_t^{(1)}$ and mean shifted by $X_t^{(2)}$. I set $\beta_0^{(1)} = \beta_0^{(2)} = 1$. I consider three different numbers of categories: $J \in \{1, 2, 3\}$. I set $(\gamma_0^1, \gamma_0^2, \gamma_0^3) = (0, 1, 2)$. In the empirical context of female labor force participation, $Y_t$ may represent different levels of labor force participation, such as not working, working part-time, or working full-time, $X_t^{(1)}$ and $X_t^{(2)}$ may represent husband's income and fertility, respectively, and $A$ may capture latent household productivity or access to job networks.

Fixing a counterfactual value $\underline{x} = (-0.5, 1)$ for $X_t$, the object of interest is the counterfactual survival probability $\Pr(Y_t(\underline{x}) \geq 1)$. I compute the sharp bounds on $\Pr(Y_t(\underline{x}) \geq 1)$ using Theorem 2 and noting that

$$\Pr(Y_t(\underline{x}) \geq 1) = \int F_{Y_t(\underline{x})|X=x}([1, \infty))dF_X(x),$$

where the integral is approximated by 5,000 random draws. Figure 4 shows the sharp bounds on $\Pr(Y_t(\underline{x}) = 1)$ re-centered by the true value for $J \in \{1, 2, 3\}$ and $T \in \{1, 2, \ldots, 20\}$. One can see that the bounds tighten as $T$ increases. There are substantial gains in identifying power when $T$ increases from 1 to 10, but the incremental gains are less pronounced when $T$ further increases from 10 to 20. The bound widths do not differ much across $J$, especially when $T$ is relatively large.

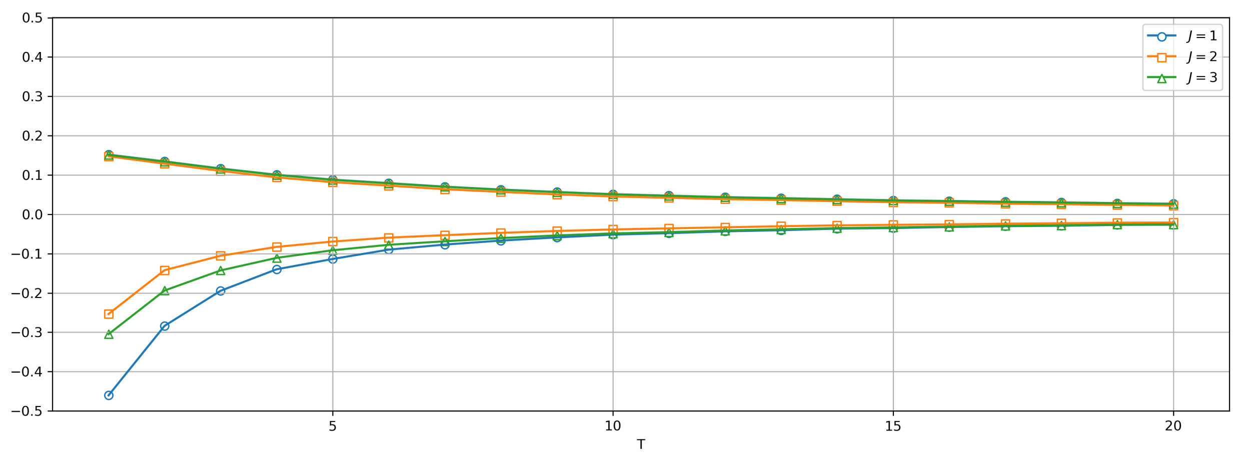



Figure 4: Re-centered Sharp Bounds on $\Pr(Y_t(\underline{x}) \geq 1)$ in Ordered Choice Models

For Example 4, I consider the following data generating process:

$$Y_t = \arg\max_j Y^*_{jt}, \ t = 1, \ldots, T,$$

where the indirect utilities are given by

$$\begin{aligned} Y^*_{0t} &= 0, \\ Y^*_{jt} &= \beta_0^{(1)} X_{jt}^{(1)} + \beta_0^{(2)} X_{jt}^{(2)} + U_{jt}, \ j = 1, \ldots, J. \end{aligned}$$

Similar to Example 2, $X^{(1)}_{jt} \sim N(0, 0.5)$ $\forall j$ and $U_{jt} = A_j + V_{jt}$ $\forall j$, where $(V_{1t}, \ldots, V_{Jt})$ follows a zero mean multivariate normal distribution with a variance matrix that has 0.5 on the diagonal and 0.25 in all off-diagonal elements. I define two equally sized latent populations of cross-sectional units. In the first population, $X_{jt}^{(2)} \sim \text{Bernoulli}(0.5)\ \forall j$ and $A_j = 1 + (0.5 + T(\bar{X}_j^{(1)})^2) \cdot Z_j\ \forall j$, while in the second population, $X_{jt}^{(2)} = 0\ \forall j$ and $A_j = (0.5 + T(\bar{X}_j^{(1)})^2) \cdot Z_j\ \forall j$, where $\bar{X}_j^{(1)} = \frac{1}{T}\sum_{t=1}^{T} X_{jt}^{(1)}$ and $Z_1, \ldots, Z_J$ are independent $N(0, 0.5)$ random variables. Here again, $A_j$ exhibits heteroskedasticity driven by $X_{jt}^{(1)}$ and a shift in mean based on $X_{jt}^{(2)}$. I set $\beta_0^{(1)} = \beta_0^{(2)} = 1$. I consider three different numbers of alternatives: $J \in \{1, 2, 3\}$. In the empirical context of consumers choosing among different brands, $X_{jt}^{(1)}$ may represent prices, $X_{jt}^{(2)}$ may represent promotion status, and $A_j$ may capture quality and intrinsic brand preference.

Fixing counterfactual values $\underline{x}_1 = (-0.5, 1)$ for $X_{1t}$ and $\underline{x}_j = (0, 0)$ for $X_{jt}\ \forall j > 1$, the object of interest is the probability of alternative 1 being chosen: $\Pr(Y_t(\underline{x}) = 1)$. I compute the sharp bounds on $\Pr(Y_t(\underline{x}) = 1)$ using Theorem 3 and noting that

$$\Pr(Y_t(\underline{x}) = 1) = \int F_{Y_t(\underline{x})|X=x}(\{1\}) dF_X(x),$$

where the integral is approximated by 5,000 random draws. Figure 5 shows the sharp bounds on $\Pr(Y_t(\underline{x}) = 1)$ re-centered by the true value for $J \in \{1, 2, 3\}$ and $T \in \{1, 2, \ldots, 20\}$. The trend in identifying power as T increases aligns with the pattern observed in Figure 4. Unlike in Figure 4, the bounds become wider when $J$ increases. A plausible explanation is that unlike Example 2, here a larger $J$ leads to higher-dimensional unobserved heterogeneity, whose distribution may require more data to learn about.

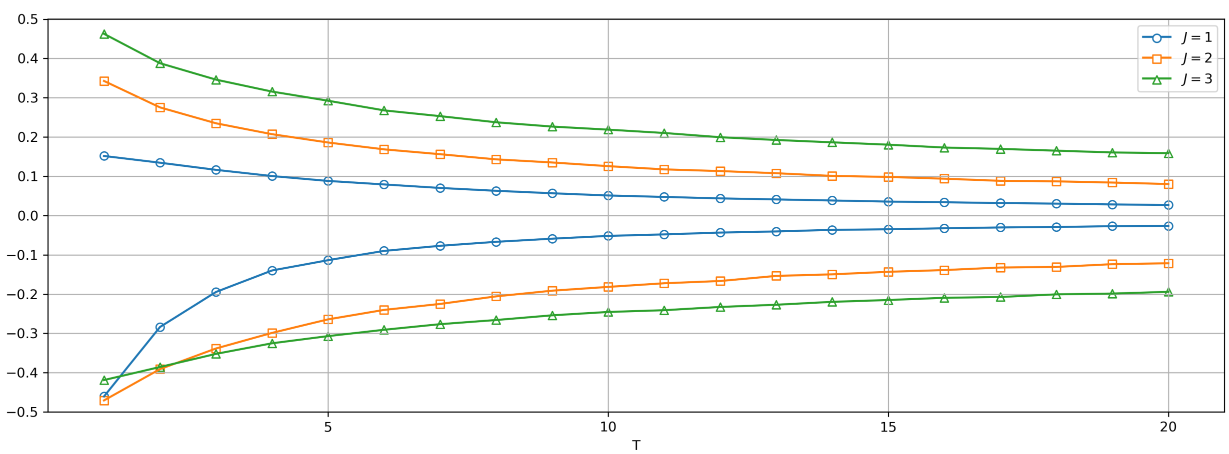



Figure 5: Re-centered Sharp Bounds on $\Pr(Y_t(\underline{x}) = 1)$ in Multinomial Choice Models

# 7 Empirical Applications

## 7.1 Binary Choice Model: Female Labor Force Participation

In the first empirical illustration, I study women's labor force participation using data from the US Panel Study of Income Dynamics (PSID) and the British Household Panel Survey (BHPS). For the PSID, I use a sample from Fernández-Val (2009), which consists of $N = 1461$ women over $T = 9$ years between 1980-1988. Only married women aged 18-64 with husbands in the labor force in each sample period are included. For the BHPS, I construct a similar sample from all 1991-2008 waves, which consists of $N = 4602$ women. The sample is an unbalanced panel, in which any woman observed in at least two waves is included.

For illustrative purposes, I focus on the static binary choice model:

$$Y_{it} = 1\{X_{it}^\top \beta_0 + U_{it} \geq 0\},$$

where $Y_t$ is the labor force participation indicator, and $X_t$ includes the natural logarithm of the husband's income, the number of children in three age categories, and a quadratic function of age. Note that some unobserved factors, such as household productivity and access to job networks, may simultaneously affect both fertility and a husband's income, as well as labor force participation. I assume that these factors are invariant over time so that Assumption 1 holds. I interpret the age categories in the two samples as follows: the PSID divides children into infants (0-2 years), preschoolers (3-5 years), and school-age children (6-17 years), while the BHPS divides children into infants (0-2 years), preschoolers (3-4 years), and school-age children (5-18 years). Descriptive statistics for both samples are given in Table 1.

Table 1: Descriptive Statistics

| | PSID Sample | | BHPS Sample | |
|---|---|---|---|---|
| | Mean | Std. Dev. | Mean | Std. Dev. |
| Participation | 0.72 | 0.45 | 0.78 | 0.41 |
| Age | 37.3 | 9.22 | 41.9 | 10.02 |
| Infants | 0.23 | 0.47 | 0.12 | 0.34 |
| Preschoolers | 0.29 | 0.51 | 0.12 | 0.34 |
| School-Age Children | 1.05 | 1.10 | 0.74 | 0.98 |
| Husband's Income (1995 \$1000/£1000) | 42.29 | 40.01 | 20.02 | 15.46 |
| No. Observations | 13149 | | 35608 | |

Continuous variation in the husband's income enables the point identification of $\beta_0$. I estimate $\beta_0$ using the maximum-score-type objective function:

$$\sum_i \sum_{t>s} (Y_{it} - Y_{is}) \cdot \operatorname{sgn}((X_{it} - X_{is})^\top \beta).$$

Table 2 reports the point estimates of $\beta_0$. One can see that the coefficients on the number of children in all three age categories are consistent across samples, exhibiting the same sign and similar magnitudes. While the coefficients on log husband's income also have the same sign in both samples, the magnitude is notably smaller in the BHPS sample. The coefficients on age and age squared indicate a concave relationship.

Table 2: Estimated $\beta_0$

| | PSID Sample | | | BHPS Sample | |
|---|---|---|---|---|---|
| | Max. Score | Pooled Logit | FE Logit | Max. Score | Pooled Logit |
| Infants | -1 | -1 | -1 | -1 | -1 |
| Preschoolers | -0.57 | -0.60 | -0.58 | -0.60 | -0.69 |
| School-Age Children | -0.01 | -0.17 | -0.19 | -0.01 | -0.27 |
| Log Husband's Income | -0.10 | -0.38 | -0.34 | -0.01 | -0.06 |
| Age/10 | 1.14 | 1.74 | 3.35 | 1.02 | 2.16 |
| $(\text{Age}/10)^2$ | -0.13 | -0.27 | -0.42 | -0.11 | -0.28 |

Consider the counterfactual scenario where log husband's income and age are set at their time averages and the number of children in each age category is increased from 0 to 1. I calculate the sharp bounds on counterfactual probabilities of labor force participation using the estimator developed in Section 5 and plot them in Figure 6. To do this, I plug in the maximum-score estimates of $\beta_0$ in Table 2 and the estimates of observed conditional choice probabilities, $\tau_0(x)$, from the logistic regression of observed choices on $X_{it}$ and $\frac{1}{T_i}\sum_{t=1}^{T_i} X_{it}$.[4]

In both samples, the bounds predict a decrease in the labor force participation rate when having one more infant or preschooler, while the effect of having one more school-age child is ambiguous. On the other hand, the bounds for having one infant or preschooler are wider than those for having one school-age child. One plausible explanation is that over 91% of the observations in the PSID sample and over 96% in the BHPS sample have either no infant or no preschooler. These observations tend to have a higher index compared to the counterfactual, providing an informative upper bound and

[4]Note that under Assumption 1, each element of $\tau_0(x)$ can be written as $F_{Y_t|X=x}(\{y\}) = F_{U_t|X=x}(\mathcal{U}(y, x_t; \theta_0)) = G(x_t, x)$. Hence, the logistic regression of observed choices on some function of $X_{it}$ and lower-dimensional statistics of $X_i$, such as $\frac{1}{T_i}\sum_{t=1}^{T_i} X_{it}$, can be viewed as a series logit approximation to $\tau_0(x)$.

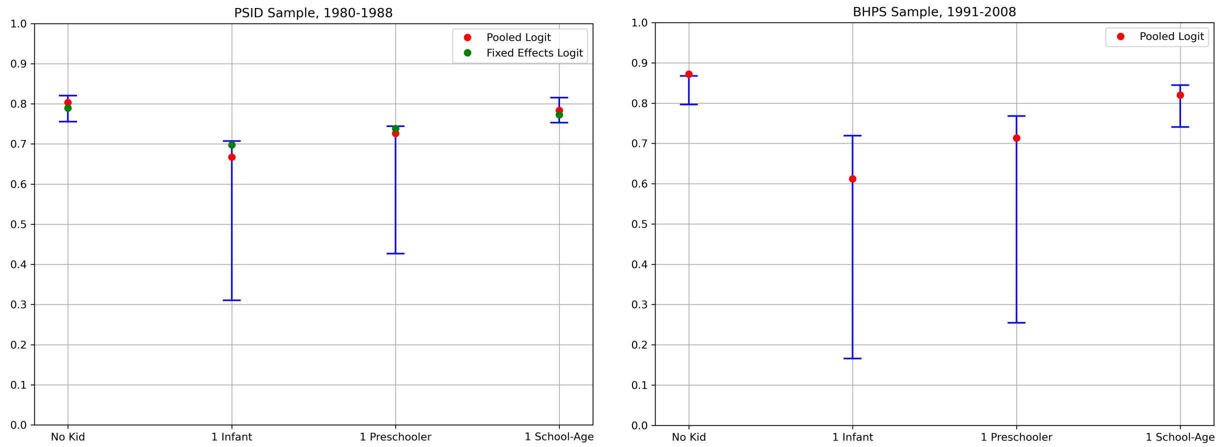



Figure 6: Counterfactual Probabilities of Labor Force Participation

a trivial lower bound. Overall, there seems to be a common pattern in how fertility affects female labor force participation across different countries and non-overlapping time periods.

For comparison, I also consider two parametric specifications. I assume $U_{it} = A_i + V_{it}$, where $V_{it} \perp X_{it}$ and $V_{it}$ is distributed i.i.d. Type 1 extreme value. In the first specification "Pooled Logit", I set $A_i = A \; \forall i$. This specification imposes exogeneity of $X_{it}$ and permits a pooled logistic regression. In the second specification "FE Logit", I do not restrict $A_i$. I first estimate $\beta_0$ using the conditional maximum likelihood estimator and then calculate the outer bound estimators for counterfactual probabilities proposed in Pakel and Weidner (2024). This specification is only applied to the PSID sample, where the panel is balanced. The associated coefficient estimates are reported in Table 2 under the columns "Pooled Logit" and "FE Logit". I plot predictions of counterfactual labor force participation rates from these two parametric specifications in Figure 6.[5] One can see that some parametric predictions lie close to the upper bounds, suggesting that they may be overly optimistic.

## 7.2 Multinomial Choice Model: Saltine Cracker Purchases

In the second empirical illustration, I apply my approach to the optical-scanner panel data set on purchases of saltine crackers in the Rome, Georgia market, collected by Information Resources Incorporated. The data set contains information on 3292 purchases of crackers by 136 households over a period of 2 years. There are three major national brands in the database: Nabisco, Sunshine, Keebler. Local brands are aggregated under the "Private" label. The data set also includes three

[5]The bounds based on FE Logit are quite tight, with widths smaller than $10^{-4}$, so I only report the midpoints.

explanatory variables, two of which are binary, and the other one is continuous. The first binary explanatory variable, "display", denotes whether or not a brand was on special display at the store at the time of purchase. The second binary explanatory variable, "feature", denotes whether or not a brand was featured in a newspaper advertisement at the time of purchase. The third explanatory variable is the "price", which corresponds to the actual price (in dollars) for the brand purchased and the shelf price for all other brands. Table 3 reports the descriptive statistics for each brand.

Table 3: Data Characteristics of Saltine Crackers

| | Nabisco | Sunshine | Keebler | Private |
|---|---|---|---|---|
| Market Share | 0.54 | 0.07 | 0.07 | 0.32 |
| Display | 0.34 | 0.13 | 0.11 | 0.10 |
| Feature | 0.09 | 0.04 | 0.04 | 0.05 |
| Average Price | 1.08 | 0.96 | 1.13 | 0.68 |

The dataset is an unbalanced panel data with the number of purchases varying across households $i$ $(\equiv T_i, 14 \leq T_i \leq 77)$. Write $\bar{\mathcal{J}} = \{1 = \text{Nabisco}, 2 = \text{Sunshine}, 3 = \text{Keebler}, 4 = \text{Private}\}$ for the choice set. For each household $i$, brand $j$, and purchase $t$, I use $X_{ijt}^{(1)}$, $X_{ijt}^{(2)}$, and $X_{ijt}^{(3)}$ to denote the three explanatory variables: the logarithm of "price", "display", and "feature", respectively. There are unobserved confounders, such as quality and intrinsic brand preferences, which are likely to remain invariant during the sample period. Hence, Assumption 1 is plausibly valid.

I follow Khan et al. (2021) to model the observed choice as

$$Y_{ijt} = 1\{Y_{ijt}^* > Y_{ikt}^*,\ \forall k \neq j\},$$

where the indirect utilities are given by

$$Y_{ijt}^* = -X_{ijt}^{(1)} + \beta_0^{(1)} X_{ijt}^{(2)} + \beta_0^{(2)} X_{ijt}^{(3)} + U_{ijt}, \quad j \in \bar{\mathcal{J}}, t = 1, \ldots, T_i,$$

where the coefficient on $X_{ijt}^{(1)}$ is normalized to be $-1$. $(\beta_0^{(1)}, \beta_0^{(2)})$ is point-identified because of rich variation in prices and can be estimated by minimizing a localized rank-based objective function

$$\sum_i \sum_{t>s} K_{h_n}(X_{i(-1)s}^{(1)} - X_{i(-1)t}^{(1)}) 1\{\tilde{X}_{i(-1)s} = \tilde{X}_{i(-1)t}\}(Y_{i1s} - Y_{i1t}) \cdot \text{sgn}((X_{i1s} - X_{i1t})^\top \beta),$$

where $\beta = (-1, \beta^{(1)}, \beta^{(2)})^\top$, $\tilde{X}_{ijt} = (X_{ijt}^{(2)}, X_{ijt}^{(3)})'$, and $X_{i(-1)t}^{(1)}$ $(\tilde{X}_{i(-1)t})$ denotes the vector collecting

$X_{ijt}^{(1)}$ ($\tilde{X}_{ijt}$) for all $j \in \bar{\mathcal{J}} \setminus \{1\}$. Following Khan et al. (2021), I choose the Gaussian kernel function and $h_n = 3\hat{\sigma}n^{-1/6}/\sqrt[3]{\log n}$, where $\hat{\sigma}$ is the standard deviation of the matching variable.

No other methods in the literature deliver counterfactual predictions for panel multinomial choice models. For comparison, I consider two parametric models, pooled multinomial logit and pooled multinomial probit, based on the indirect utility specification

$$Y_{ijt}^{*} = -\beta_0^{(0)} X_{ijt}^{(1)} + \beta_0^{(1)} X_{ijt}^{(2)} + \beta_0^{(2)} X_{ijt}^{(3)} + \alpha_j + V_{ijt}, \quad j \in \bar{\mathcal{J}}, t = 1, \ldots, T_i,$$

where $V_{ijt}$ is independent of $X_{ijt}$, and $(\beta_0^{(0)}, \beta_0^{(1)}, \beta_0^{(2)})$ and alternative-specific intercepts $\alpha_j$ are parameters to be estimated.[6] Table 4 reports the point estimates of coefficients.[7]

Table 4: Parametric and Semiparametric Estimations of Coefficients

| | $\hat{\beta}^{(1)}$ | $\hat{\beta}^{(2)}$ |
|---|---|---|
| Semiparametric panel | 0.08 | 0.09 |
| Pooled multinomial logit | 0.03 | 0.16 |
| Pooled multinomial probit | 0.02 | 0.11 |

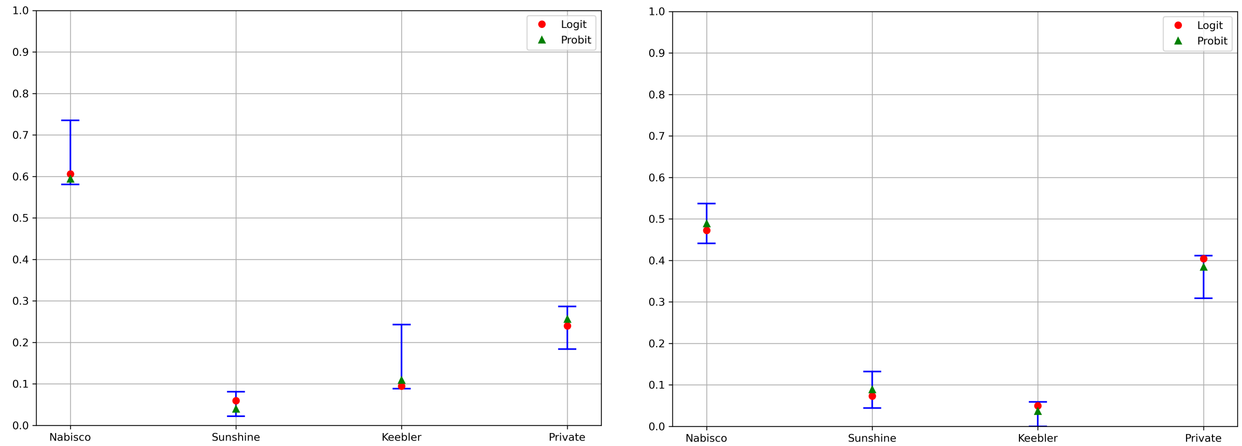
I consider the counterfactual choice probabilities under two counterfactual values $\underline{x}$ and $\overline{x}$ for explanatory variables. The price vector for $\underline{x}$ is $\underline{p} = (1.09, 1.05, 1.05, 0.78)$ and the price vector for $\overline{x}$ is $\overline{p} = (1.09, 0.89, 1.21, 0.59)$. The display and feature statuses are fixed at zero for all brands for both $\underline{x}$ and $\overline{x}$. Moving from $\underline{x}$ to $\overline{x}$ corresponds to a simultaneous price change of multiple brands, which consists of a rise from the 25th percentile to the 75th percentile of the price for brand 3 (Keebler), and a fall from the 75th percentile to the 25th percentile of the price for brands 2 and 4 (Sunshine and Private), with the price of brand 1 (Nabisco) fixed at the median.

I calculate the sharp bounds on counterfactual choice probabilities using the estimator developed in Section 5. To do this, I plug in the semiparametric estimates of $(\beta_0^{(1)}, \beta_0^{(2)})$ in Table 4 and the estimates of observed conditional choice probabilities, $\tau_0(x)$, from multinomial logistic regression of observed choices on $\{(X_{ijt}, (X_{ijt}^{(1)})^2), \frac{1}{T_i}\sum_{t=1}^{T_i} X_{ijt}, \frac{1}{T_i}\sum_{t=1}^{T_i}(X_{ijt}^{(1)})^2)\}_{j \in \bar{\mathcal{J}}}$. Panels (a) and (b) of Figure 7 display the bounds under $\underline{x}$ and $\overline{x}$, respectively.

The bounds predict a market share decrease for brands 1 and 3 (Nabisco and Keebler) and a market share increase for brand 4 (Private), while the direction of the market share change for brand

[6]The parameter estimation of these models is conducted using Stata packages "cmclogit" and "cmcmmprobit".

[7]For the pooled multinomial logit and probit models, I report the ratios of the coefficients on $X_{ijt}^{(2)}$ and $X_{ijt}^{(3)}$ to the absolute value of the coefficient on $X_{ijt}^{(1)}$.




(a) Under Price Vector $\underline{p} = (1.09, 1.05, 1.05, 0.78)$ (b) Under Price Vector $\overline{p} = (1.09, 0.89, 1.21, 0.59)$

Figure 7: Counterfactual Choice Probabilities

2 (Sunshine) is ambiguous. For comparison, I also plot the predictions from pooled multinomial logit and probit models in Figure 7. Parametric predictions lie within semiparametric bounds, with some close to upper or lower limits. Consequently, parametric models might underestimate the market share change of brand 3 (Keebler).

# 8 Extension: Dynamic Binary Choice Models

Although the main framework of this paper focuses on static models, the identification strategy based on the set inclusion relationship of $U$-level sets can be applied to dynamic models to derive (non-sharp) identifying restrictions on counterfactual distributions. To demonstrate this, I consider the dynamic panel data binary choice model:

$$Y_t = 1\{\rho_0 Y_{t-1} + X_t^\top \beta_0 + U_t \geq 0\}.$$

Let $\theta_0 = (\rho_0, \beta_0)$. I maintain Assumption 1, which is termed *partial stationarity* in Gao and Wang (2024) because the conditioning set only contains part of the explanatory variables. Identification of $\theta_0$ under Assumption 1 is discussed in Khan et al. (2023) and Gao and Wang (2024). Fixing a counterfactual value $(\underline{y}, \underline{x})$ for $(Y_{t-1}, X_t)$, the interest is in the distribution of the counterfactual outcome $Y_t(\underline{y}, \underline{x}) = 1\{\rho_0 \underline{y} + \underline{x}^\top \beta_0 + U_t \geq 0\}$. This is in line with the *dynamic potential outcome* model of Torgovitsky (2019).

I slightly modify the definition of $U$-level sets as

$$\mathcal{U}(y_t, y_{t-1}, x_t; \theta) = \{u_t : y_t = 1\{\rho y_{t-1} + x_t^\top \beta + u_t \geq 0\}\}.$$

The key observation is that for $y \in \{0,1\}$,

$$U_t \in \mathcal{U}(y, Y_{t-1}, X_t; \theta_0) \text{ and } \mathcal{U}(y, Y_{t-1}, X_t; \theta_0) \subseteq \mathcal{U}(y, \underline{y}, \underline{x}; \theta_0) \;\Rightarrow\; U_t \in \mathcal{U}(y, \underline{y}, \underline{x}; \theta_0). \tag{10}$$

Note that

$$\begin{aligned}
& \mathcal{U}(1, Y_{t-1}, X_t; \theta_0) \subseteq \mathcal{U}(1, \underline{y}, \underline{x}; \theta_0) \\
\iff \quad & \mathcal{U}(0, Y_{t-1}, X_t; \theta_0) \supseteq \mathcal{U}(0, \underline{y}, \underline{x}; \theta_0) \\
\iff \quad & (Y_{t-1} = 1 \text{ and } \rho_0 \underline{y} + \underline{x}^\top \beta_0 \geq \rho_0 + X_t^\top \beta_0) \text{ or } (Y_{t-1} = 0 \text{ and } \rho_0 \underline{y} + \underline{x}^\top \beta_0 \geq X_t^\top \beta_0).
\end{aligned}$$

Taking the conditional expectation of (10) given $X = x$ yields

$$B_t^l(x; \theta_0) \leq \Pr(Y_t(\underline{y}, \underline{x}) = 1 | X = x) \leq B_t^u(x; \theta_0),$$

where

$$\begin{aligned}
B_t^l(x; \theta) &= \begin{cases} \Pr(Y_t = 1 | X = x) & \text{if } \rho \underline{y} + \underline{x}^\top \beta \geq \max\{\rho + x_t^\top \beta, x_t^\top \beta\} \\ \Pr(Y_t = 1, Y_{t-1} = 0 | X = x) & \text{if } x_t^\top \beta \leq \rho \underline{y} + \underline{x}^\top \beta < \rho + x_t^\top \beta \\ \Pr(Y_t = 1, Y_{t-1} = 1 | X = x) & \text{if } \rho + x_t^\top \beta \leq \rho \underline{y} + \underline{x}^\top \beta < x_t^\top \beta \\ 0 & \text{otherwise} \end{cases}, \\
B_t^u(x; \theta) &= \begin{cases} 1 & \text{if } \rho \underline{y} + \underline{x}^\top \beta \geq \max\{\rho + x_t^\top \beta, x_t^\top \beta\} \\ 1 - \Pr(Y_t = 0, Y_{t-1} = 1 | X = x) & \text{if } x_t^\top \beta \leq \rho \underline{y} + \underline{x}^\top \beta < \rho + x_t^\top \beta \\ 1 - \Pr(Y_t = 0, Y_{t-1} = 0 | X = x) & \text{if } \rho + x_t^\top \beta \leq \rho \underline{y} + \underline{x}^\top \beta < x_t^\top \beta \\ \Pr(Y_t = 1 | X = x) & \text{otherwise} \end{cases}.
\end{aligned}$$

The intuition is that when the counterfactual index is large or small enough to eliminate uncertainty in the set inclusion relationship of $U$-level sets, the bounds align with those in the static case. Otherwise, the bounds will depend on the distribution of the lagged outcome.

Assumption 1 allows me to use information across all time periods to obtain tighter bounds. Eventually, the counterfactual probability $\Pr(Y_t(\underline{y}, \underline{x}) = 1)$ can be bounded as

$$E\Big[\max_t B_t^l(X;\theta_0)\Big] \leq \Pr(Y_t(\underline{y}, \underline{x}) = 1) \leq E\Big[\min_t B_t^u(X;\theta_0)\Big].$$

Further analysis for nonlinear dynamic panel data models is left to future research.

# 9 Conclusion

This paper establishes the sharp identified set of the distribution of the counterfactual outcome in semiparametric nonlinear panel data models in cases where structural parameters are point-identified. I rely on time homogeneity of the distribution of unobserved heterogeneity while allowing for flexible dependence between unobserved heterogeneity and explanatory variables. I provide tractable implementation procedures for monotone transformation models and multinomial choice models, by exploiting an index separability condition. I examine factors affecting the informativeness of the identified set through numerical experiments. I also derive theoretical results for estimation and inference. My approach is applied to empirical data on female labor force participation and purchases of saltine crackers. Finally, I discuss the potential extension of my identification strategy to dynamic settings.

# References


Abrevaya, J. and J. Huang (2005): "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, 73, 1175–1204.

Bhattacharya, D. (2015): "Nonparametric welfare analysis for discrete choice," *Econometrica*, 83, 617–649.

——— (2018): "Empirical welfare analysis for discrete choice: Some general results," *Quantitative Economics*, 9, 571–615.

Blundell, R. W. and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Cambridge: Cambridge University Press, vol. 2, 312–357.

——— (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, 71, 655–679.

Botosaru, I. and C. Muris (2024): "Identification of time-varying counterfactual parameters in nonlinear panel models," *Journal of Econometrics*, 105639.

Botosaru, I., C. Muris, and K. Pendakur (2023): "Identification of time-varying transformation models with fixed effects, with an application to unobserved heterogeneity in resource shares," *Journal of Econometrics*, 232, 576–597.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020): "Bootstrap-Based Inference for Cube Root Asymptotics," *Econometrica*, 88, 2203–2219.

Charlier, E., B. Melenberg, and A. H. O. van Soest (1995): "A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation," *Statistica Neerlandica*, 49, 324–342.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

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, annals Issue in Honor of Jerry A. Hausman.

Chesher, A. and A. M. Rosen (2017): "Generalized Instrumental Variable Models," *Econometrica*, 85, 959–989.

Chesher, A., A. M. Rosen, and Y. Zhang (2024): "Robust Analysis of Short Panels," ArXiv: 2401.06611.

Chiong, K. X., Y.-W. Hsieh, and M. Shum (2021): "Bounds on counterfactuals in semiparametric discrete-choice models," in *Handbook of Research Methods and Applications in Empirical Microeconomics*, Edward Elgar Publishing, 223–237.

Davezies, L., X. D'Haultfoeuille, and L. Laage (2022): "Identification and Estimation of Average Marginal Effects in Fixed Effects Logit Models," ArXiv: 2105.00879.

Delgado, M. A., J. M. Rodríguez-Poo, and M. Wolf (2001): "Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator," *Economics Letters*, 73, 241–250.

Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150, 71–85.

Gao, W. Y. and M. Li (2020): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," ArXiv: 2009.00085.

Gao, W. Y. and R. Wang (2024): "Identification of Nonlinear Dynamic Panels under Partial Stationarity," ArXiv: 2401.00264.

Gu, J., T. Russell, and T. Stringham (2024): "Counterfactual identification and latent space enumeration in discrete outcome models," Available at SSRN: https://ssrn.com/abstract=4188109 or http://dx.doi.org/10.2139/ssrn.4188109.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hoderlein, S. and H. White (2012): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics*, 168, 300–314.

Hong, H. and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48, 397 – 412.

Honoré, B. E. and E. Kyriazidou (2000): "Estimation of Tobit-Type Models with Individual Specific Effects," *Econometric Reviews*, 19, 341–66.

Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

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

Khan, S., M. Ponomareva, and E. Tamer (2016): "Identification of panel data models with endogenous censoring," *Journal of Econometrics*, 194, 57–75.

——— (2023): "Identification of dynamic binary response models," *Journal of Econometrics*, 237, 105515.

Lee, S. M. S. and M. C. Pun (2006): "On m out of n Bootstrapping for Nonstandard M-Estimation With Nuisance Parameters," *Journal of the American Statistical Association*, 101, 1185–1197.

Liu, L., A. Poirier, and J.-L. Shiu (2021): "Identification and estimation of average partial effects in semiparametric binary response panel models," *arXiv preprint arXiv:2105.12891*.

Manski, C. F. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica: Journal of the Econometric Society*, 357–362.

——— (2007): "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48, 1393–1410.

Pakel, C. and M. Weidner (2024): "Bounds on Average Effects in Discrete Choice Panel Data Models," ArXiv: 2309.09299.

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

Semenova, V. (2024): "Aggregated Intersection Bounds and Aggregated Minimax Values," ArXiv: 2303.00982.

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

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91, 107–146.

Torgovitsky, A. (2019): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505.

# Appendix A Proofs

*Proof of Theorem 1.* Following Chesher and Rosen (2017), I adopt the notion of structures. In my case, a structure is a pair $m = (\theta, \mathcal{F}_{U|X})$. Each structure $m$ delivers a conditional distribution $P_{Y|X}(\cdot|x;m)$ for each $x \in \mathrm{Supp}(X)$. Let $\mathcal{P}_{Y|X}(m) = \{P_{Y|X}(\cdot|x;m) : x \in \mathrm{Supp}(X)\}$. Let $\mathcal{M}$ be the set of structures that satisfy Assumption 1. Let $\mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ denote the set of structures identified by $\mathcal{M}$ and $\mathcal{F}_{Y|X}$, that is, $m \in \mathcal{M}$ if $m$ is admitted by $\mathcal{M}$ and $\mathcal{F}_{Y|X}$ and $\mathcal{P}_{Y|X}(m)$ agree. Then, the sharp identified set for $\mathcal{F}_{Y_t(\underline{x})|X}$ is defined as

$$\begin{aligned}\mathsf{F}^*_{Y_t(\underline{x})|X} = \{&\mathcal{F}_{Y_t(\underline{x})|X} : \exists(\theta, \mathcal{F}_{U|X}) \in \mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X}) \\ &\text{s.t. } \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta)) \ \textit{a.e. } x \in \mathrm{Supp}(X)\}.\end{aligned}$$

Note that $\mathsf{F}^*_{Y_t(\underline{x})|X}$ depends on $\mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ only through $(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T)$. Let $\mathcal{I}^*(\mathcal{M}, \mathcal{F}_{Y|X})$ denote the projection of $\mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ onto $(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T)$. Then,

$$\begin{aligned}\mathsf{F}^*_{Y_t(\underline{x})|X} = \{&\mathcal{F}_{Y_t(\underline{x})|X} : \exists(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T) \in \mathcal{I}^*(\mathcal{M}, \mathcal{F}_{Y|X}) \\ &\text{s.t. } \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta)) \ \textit{a.e. } x \in \mathrm{Supp}(X)\}. \qquad (11)\end{aligned}$$

In static models, $(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T)$ only deliver the marginals of $P_{Y|X}(\cdot|x;m)$. By Sklar's theorem, there exists a collection of $T$-variate copula $\mathcal{C}_X = \{C_X(\cdot|x) : x \in \mathrm{Supp}(X)\}$ such that $C_X(\cdot|x)$ reproduces the dependence structure of $P_{Y|X}(\cdot|x;m)$. In this sense, $(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T, \mathcal{C}_X)$ is observational equivalent to $m$. Since Assumption 1 only restricts $\{\mathcal{F}_{U_t|X}\}_{t=1}^T$, one can set $\mathcal{C}_X$ to be the collection of copulas associated with $\mathcal{F}_{Y|X}$ and require $(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T)$ to satisfy Assumption 1 and be consistent with the marginals of $\mathcal{F}_{Y|X}$. Hence,

$$\begin{aligned}\mathcal{I}^*(\mathcal{M}, \mathcal{F}_{Y|X}) = \{&(\theta, \{\mathcal{F}_{U_t|X}\}_{t=1}^T) : \text{ Assumption 1 holds and } \forall t \in \{1, \ldots, T\}, \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), \\ &F_{Y_t|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, x_t; \theta)) \ \textit{a.e. } x \in \mathrm{Supp}(X)\}.\end{aligned}$$

Finally, by Assumption 2, one can further write

$$\begin{aligned}\mathcal{I}^*(\mathcal{M}, \mathcal{F}_{Y|X}) = \{\theta_0\} \times \{&\{\mathcal{F}_{U_t|X}\}_{t=1}^T : \text{ Assumption 1 holds and } \forall t \in \{1, \ldots, T\}, \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), \\ &F_{Y_t|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, x_t; \theta_0)) \ \textit{a.e. } x \in \mathrm{Supp}(X)\}\end{aligned}$$

$$= \{\theta_0\} \times \bigcap_{t'=1}^{T} \{\mathcal{F}_{U_t|X} : \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}),$$

$$F_{Y_{t'}|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, x_{t'}; \theta_0)) \text{ } a.e. \text{ } x \in \text{Supp}(X)\}. \tag{12}$$

The result follows by plugging (12) into (11). □

*Proof of (5).* Fix $(\mathcal{T}, \mathcal{T}') \in \mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta)$. By definition, $\mathcal{T}' = \mathcal{T}$. For any $j \in \mathcal{T}$ and $k \notin \mathcal{T}$, $(x_{jt} - \underline{x}_j)^\top \beta \geq (x_{kt} - \underline{x}_k)^\top \beta$. Re-arranging, $(x_{jt} - x_{kt})^\top \beta \geq (\underline{x}_j - \underline{x}_k)^\top \beta$. Take any $U_t \in \mathcal{U}(\mathcal{T}, \underline{x}; \theta)$. Then, there exists $j \in \mathcal{T}$ such that for any $k \notin \mathcal{T}$,

$$U_{kt} - U_{jt} \leq (\underline{x}_j - \underline{x}_k)^\top \beta \leq (x_{jt} - x_{kt})^\top \beta.$$

Hence, $U_t \in \mathcal{U}(\mathcal{T}, x_t; \theta)$. □

*Proof of Theorem 2.* By definition, $\mathcal{F}_{U_t|X} \in \mathsf{F}^*_{U_t|X}$ if and only if $\forall y' \in \mathcal{Y}$, $\forall t' \in \{1, \ldots, T\}$,

$$F_{Y_{t'}|X=x}([y', \infty)) = F_{U_t|X=x}(\mathcal{U}([y', \infty), x_{t'}; \theta_0)) \text{ } a.e. \text{ } x \in \text{Supp}(X).$$

It follows that

$$\begin{aligned}
\mathsf{F}^*_{Y_t(\underline{x})|X} = \{&\mathcal{F}_{Y_t(\underline{x})|X} : \exists \mathcal{F}_{U_t|X} \text{ s.t. } \forall y \in \mathcal{Y}, \forall t' \in \{1, \ldots, T\}, \\
&F_{Y_t(\underline{x})|X=x}([y, \infty)) = F_{U_t|X=x}(\mathcal{U}([y, \infty), \underline{x}; \theta_0)), \\
&F_{Y_{t'}|X=x}([y, \infty)) = F_{U_t|X=x}(\mathcal{U}([y, \infty), x_{t'}; \theta_0)) \text{ } a.e. \text{ } x \in \text{Supp}(X)\} \\
= \{&\mathcal{F}_{Y_t(\underline{x})|X} : \exists \mathcal{F}_{U_t|X} \text{ s.t. } \forall y \in \mathcal{Y}, \forall t' \in \{1, \ldots, T\}, \\
&F_{Y_t(\underline{x})|X=x}([y, \infty)) = F_{U_t|X=x}([-\underline{x}^\top \beta_0 + h^-(y, \gamma_0), \infty)), \\
&F_{Y_{t'}|X=x}([y, \infty)) = F_{U_t|X=x}([-x_{t'}^\top \beta_0 + h^-(y, \gamma_0), \infty)) \text{ } a.e. \text{ } x \in \text{Supp}(X)\},
\end{aligned}$$

where the second equality follows from (3). Taking $\mathcal{F}_{Y_t(\underline{x})|X}$ from the right-hand side of (6), I want to show that $\mathcal{F}_{Y_t(\underline{x})|X} \in \mathsf{F}^*_{Y_t(\underline{x})|X}$, which amounts to for all $x \in \text{Supp}(X)$ exhibiting $F_{U_t|X=x}$ satisfying $\forall y \in \mathcal{Y}$,

$$\begin{aligned}
F_{Y_t(\underline{x})|X=x}([y, \infty)) &= F_{U_t|X=x}([-\underline{x}^\top \beta_0 + h^-(y, \gamma_0), \infty)), \\
F_{Y_{t'}|X=x}([y, \infty)) &= F_{U_t|X=x}([-x_{t'}^\top \beta_0 + h^-(y, \gamma_0), \infty)), \; t' = 1, \ldots, T.
\end{aligned}$$

Fix $x \in \mathrm{Supp}(X)$. The desired $F_{U_t|X=x}$ can be constructed as follows. Define

$$\begin{aligned}
p_{t'}(y) &= F_{Y_{t'}|X=x}([y,\infty)),\ t'=1,\dots,T,\\
p_{T+1}(y) &= F_{Y_t(\underline{x})|X=x}([y,\infty)),\\
\underline{u}_{t'}(y) &= -x_{t'}^\top\beta_0 + h^-(y,\gamma_0),\ t'=1,\dots,T,\\
\underline{u}_{T+1}(y) &= -\underline{x}^\top\beta_0 + h^-(y,\gamma_0).
\end{aligned}$$

Then, (6) ensures that for any $t' \in \{1,\dots,T\}$ and $y, y' \in \mathcal{Y}$,

$$\begin{aligned}
\underline{u}_{T+1}(y) \geq \underline{u}_{t'}(y') &\iff p_{T+1}(y) \leq p_{t'}(y'),\\
\underline{u}_{T+1}(y) \leq \underline{u}_{t'}(y') &\iff p_{T+1}(y) \geq p_{t'}(y'),
\end{aligned}$$

Also, by Lemma 1 of Botosaru et al. (2023), Assumption 2 ensures that for any $t', t'' \in \{1,\dots,T\}$ and $y, y' \in \mathcal{Y}$,

$$\underline{u}_{t'}(y) \leq \underline{u}_{t''}(y') \iff p_{t'}(y) \geq p_{t''}(y').$$

Put together, for any $t', t'' \in \{1,\dots,T+1\}$ and $y, y' \in \mathcal{Y}$,

$$\underline{u}_{t'}(y) \leq \underline{u}_{t''}(y') \iff p_{t'}(y) \geq p_{t''}(y'). \tag{13}$$

For $u \in \mathbb{R}$, define

$$(t^*(u), y^*(u)) = \underset{(t',y)\in\{1,\dots,T+1\}\times\mathcal{Y}:\underline{u}_{t'}(y)\leq u}{\arg\max} \underline{u}_{t'}(y).$$

One can set

$$F_{U_t|X=x}([u,\infty)) = p_{t^*(u)}(y^*(u)),\ u \in \mathbb{R}.$$

I now show that $F_{U_t|X=x}$ satisfies the monotonicity requirement of a CDF, i.e.,

$$F_{U_t|X=x}([u,\infty)) \geq F_{U_t|X=x}([u',\infty)),\ \forall u \leq u'.$$

To see this, note that by definition,

$$\underline{u}_{t^*(u)}(y^*(u)) \leq \underline{u}_{t^*(u')}(y^*(u')).$$

which implies that

$$F_{U_t|X=x}([u,\infty)) = p_{t^*(u)}(y^*(u)) \geq p_{t^*(u')}(y^*(u')) = F_{U_t|X=x}([u',\infty)),$$

where the inequality follows from (13). $\square$

*Proof of Theorem 3.* Taking $\mathcal{F}_{Y_t(\underline{x})|X}$ from the right-hand side of (7), I want to show that $\mathcal{F}_{Y_t(\underline{x})|X} \in \mathsf{F}^*_{Y_t(\underline{x})|X}$, which amounts to for all $x \in \text{Supp}(X)$ exhibiting $F_{U_t|X=x}$ satisfying

$$\begin{aligned} F_{Y_t(\underline{x})|X=x}(\{j\}) &= F_{U_t|X=x}(\mathcal{U}(j,\underline{x};\theta_0)), \\ F_{Y_{t'}|X=x}(\{j\}) &= F_{U_t|X=x}(\mathcal{U}(j,x_{t'};\theta_0)), \end{aligned}$$

for all $j \in \{0,1,\ldots,J\}$ and $t' \in \{1,\ldots,T\}$. Fix $x \in \text{Supp}(X)$. Define $\mathcal{U}_{j_1,\ldots,j_T,j'} = \mathcal{U}(j_1,x_1;\theta_0) \cap \cdots \cap \mathcal{U}(j_T,x_T;\theta_0) \cap \mathcal{U}(j',\underline{x};\theta_0)$ and $q_{j_1,\ldots,j_T,j'} = F_{U_t|X=x}(\mathcal{U}_{j_1,\ldots,j_T,j'})$. Note that $q_{j_1,\ldots,j_T,j'} = 0$ if $\mathcal{U}_{j_1,\ldots,j_T,j'} = \emptyset$. The probabilities $q = \{q_{j_1,\ldots,j_T,j'} : \mathcal{U}_{j_1,\ldots,j_T,j'} \neq \emptyset\}$ are the building blocks for constructing $F_{U_t|X=x}$. The task can be rephrased as exhibiting $q_{j_1,\ldots,j_T,j'} \geq 0$ satisfying

$$\sum_{(j_1,\ldots,j_T,j'):\ \mathcal{U}_{j_1,\ldots,j_T,j'}\neq\emptyset,\ j'=j} q_{j_1,\ldots,j_T,j'} = F_{Y_t(\underline{x})|X=x}(\{j\}), \tag{14}$$

$$\sum_{(j_1,\ldots,j_T,j'):\ \mathcal{U}_{j_1,\ldots,j_T,j'}\neq\emptyset,\ j_{t'}=j} q_{j_1,\ldots,j_T,j'} = F_{Y_{t'}|X=x}(\{j\}), \tag{15}$$

for all $j \in \{0,1,\ldots,J\}$ and $t' \in \{1,\ldots,T\}$. Let

$$p^{\text{ct}} = \begin{bmatrix} F_{Y_t(\underline{x})|X=x}(\{0\}) \\ F_{Y_t(\underline{x})|X=x}(\{1\}) \\ \vdots \\ F_{Y_t(\underline{x})|X=x}(\{J\}) \end{bmatrix} \text{ and } p^{\text{ob}}_{t'} = \begin{bmatrix} F_{Y_{t'}|X=x}(\{0\}) \\ F_{Y_{t'}|X=x}(\{1\}) \\ \vdots \\ F_{Y_{t'}|X=x}(\{J\}) \end{bmatrix},\ t' = 1,\ldots,T.$$

Let $Q^{\text{ct}}$ be the matrix with elements in $\{0,1\}$ such that (14) can be restated as $Q^{\text{ct}}q = p^{\text{ct}}$ and let $Q^{\text{ob}}_{t'}$ be the matrix with elements in $\{0,1\}$ such that (15) can be restated as $Q^{\text{ob}}_{t'}q = p^{\text{ob}}_{t'}$. The task can be summarized as showing that $\exists q \geq 0$ such that: (A) $Q^{\text{ct}}q = p^{\text{ct}}$ and (B) $Q^{\text{ob}}_{t'}q = p^{\text{ob}}_{t'}$, $\forall t'$. Let $\{z^{t'} = (z^{t'}_0, z^{t'}_1, \ldots, z^{t'}_J)^\mathsf{T}\}_{t'=1}^T$ and $w = (w_0, w_1, \ldots, w_J)^\mathsf{T}$ be $(J+1)$-dimensional constant vectors.

Farkas's Lemma states that if

$$w^\intercal Q^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal Q^{\mathrm{ob}}_{t'} \geq 0 \text{ implies } w^\intercal p^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal p^{\mathrm{ob}}_{t'} \geq 0,$$

then $\exists q \geq 0$ satisfying constraints (A) and (B). For each $t' \in \{1, \ldots, T\}$, there exists a weak ordering for $\{(x_{jt'} - \underline{x}_j)^\top \beta_0\}_{j=0}^{J}$. Let $M_{t'}(j)$ denote the rank of alternative $j$ in this ordering and $M_{t'}^{-1}$ denote the inverse mapping. Then, $(\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}, \{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}) \in \mathbb{Y}(\underline{x}^\top \beta_0, x_{t'}^\top \beta_0)$ for $j > 0$. For any $\{a_j^{t'}\}_{j=0,1,\ldots,J,t'=1,\ldots,T} \in \mathbb{R}$,

$$\begin{aligned}
& w^\intercal p^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal p^{\mathrm{ob}}_{t'} \\
= \; & \sum_{j=0}^{J} w_j F_{Y_t(\underline{x})|X=x}(\{j\}) + \sum_{t'=1}^{T} \sum_{j=0}^{J} z_j^{t'} F_{Y_{t'}|X=x}(\{j\}) \\
= \; & \sum_{t'=1}^{T} \sum_{j=0}^{J} a^{t'}_{M_{t'}^{-1}(j)} \underbrace{(F_{Y_{t'}|X=x}(\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}) - F_{Y_t(\underline{x})|X=x}(\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}))}_{\geq 0 \text{ by } (7)} \\
& + \sum_{j=0}^{J} \Big( w_j + \sum_{t'=1}^{T} \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \Big) F_{Y_t(\underline{x})|X=x}(\{j\}) + \sum_{t'=1}^{T} \sum_{j=0}^{J} \Big( z_j^{t'} - \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \Big) F_{Y_{t'}|X=x}(\{j\}).
\end{aligned}$$

Therefore, given $w^\intercal Q^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal Q^{\mathrm{ob}}_{t'} \geq 0$, $w^\intercal p^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal p^{\mathrm{ob}}_{t'} \geq 0$ if there exist $\{a_j^{t'}\}_{j=0,1,\ldots,J,t'=1,\ldots,T} \in \mathbb{R}$ satisfying

$$\begin{aligned}
& w_j + \sum_{t'=1}^{T} \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \geq 0, \ \forall j, \\
& z_j^{t'} - \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \geq 0, \ \forall j, t', \\
& a_j^{t'} \geq 0 \text{ if } M_{t'}(j) > 0, \ \forall t'.
\end{aligned}$$

From the examination of matrices $Q^{\mathrm{ct}}$ and $Q_1^{\mathrm{ob}}, \ldots, Q_T^{\mathrm{ob}}$, $w^\intercal Q^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\intercal Q^{\mathrm{ob}}_{t'} \geq 0$ yields

$$w_{j'} + \sum_{t'=1}^{T} z^{t'}_{j_{t'}} \geq 0 \text{ if } \mathcal{U}_{j_1,\ldots,j_T,j'} \neq \emptyset.$$

For $j = 0, 1, \dots, J$, let

$$\begin{aligned}
\underline{a}_j^1 &= \min_{\ell:\, \mathcal{U}_{\ell, j_2, \dots, j_T, j} \neq \emptyset} z_\ell^1, \\
\underline{a}_j^{t'} &= \min_{\ell:\, \mathcal{U}_{\dots, j_{t'-1}, \ell, j_{t'+1}, \dots, j} \neq \emptyset} z_\ell^{t'},\ 1 < t' < T, \\
\underline{a}_j^T &= \min_{\ell:\, \mathcal{U}_{j_1, \dots, j_{T-1}, \ell, j} \neq \emptyset} z_\ell^T.
\end{aligned}$$

Then, $w_j + \sum_{t'=1}^T \underline{a}_j^{t'} \geq 0,\ \forall j$. Also, since $\mathcal{U}_{j_1, \dots, j_T, j'} \neq \emptyset$ when $j_1 = \cdots = j_T = j'$, $\underline{a}_j^{t'} \leq z_j^{t'},\ \forall j, t'$. Moreover, note that $\mathcal{U}_{j_1, \dots, j_T, j'} \neq \emptyset$ implies that $M_{t'}(j_{t'}) \geq M_{t'}(j')$, $\forall t'$. Hence, $\underline{a}^{t'}_{M_{t'}^{-1}(j)}$ is increasing in $j$. The desired $\{a_j^{t'}\}_{j=0,1,\dots,J, t'=1,\dots,T}$ can be constructed as follows:

$$\begin{aligned}
a^{t'}_{M_{t'}^{-1}(0)} &= \underline{a}^{t'}_{M_{t'}^{-1}(0)}, \\
a^{t'}_{M_{t'}^{-1}(j)} &= \underline{a}^{t'}_{M_{t'}^{-1}(j)} - \underline{a}^{t'}_{M_{t'}^{-1}(j-1)},\ j = 1, \dots, J.
\end{aligned}$$

It remains to construct $F_{U_t|X=x}$. For each $\mathcal{U}_{j_1, \dots, j_T, j'} \neq \emptyset$, choose a point $r_{j_1, \dots, j_T, j'} \in \mathcal{U}_{j_1, \dots, j_T, j'}$. Then, define $F_{U_t|X=x}$ to be the discrete distribution on support points $r_{j_1, \dots, j_T, j'}$ with $F_{U_t|X=x}(\{r_{j_1, \dots, j_T, j'}\}) = q_{j_1, \dots, j_T, j'}$. Now it can be concluded that (7) holds. $\square$

*Proof of Theorem 4.* By noting that

$$\arg\max_{\lambda \in \Lambda_l(x;\theta)} \lambda^\top \tau(x) = -\arg\min_{\lambda \in \Lambda_l(x;\theta)} -\lambda^\top \tau(x),$$

it suffices to focus on the upper bound. Henceforth, I suppress the $u$ subscript for ease of notation. For each function $f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, let $\mathbb{G}_N(f(Y, X)) = N^{-1/2} \sum_{i=1}^N (f(Y_i, X_i) - E[f(Y_i, X_i)])$. The standard decomposition gives

$$\begin{aligned}
\sqrt{N}(\hat{\Psi}(\theta) - \Psi(\theta)) = \mathbb{G}_n\Big( & \sum_{\lambda \in \Lambda(X;\theta)} 1\{\lambda^*(X; \theta, \tau_0) = \lambda\} \lambda^\top I(Y) \Big) && (16) \\
&+ \mathbb{G}_n\Big( \sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X; \theta, \hat{\tau}) = \lambda\} - 1\{\lambda^*(X; \theta, \tau_0) = \lambda\}) \lambda^\top I(Y) \Big) && (17) \\
&+ \sqrt{N} E\Big[ \sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X; \theta, \hat{\tau}) = \lambda\} - 1\{\lambda^*(X; \theta, \tau_0) = \lambda\}) \lambda^\top I(Y) \Big]. && (18)
\end{aligned}$$

To show (17) and (18) are $o_p(1)$, I will use the following lemma:

**Lemma 1.** *Suppose that Assumptions 3 and 5 hold. Then, for all $\theta$, there exists $C > 0$ such that for any $\delta \geq 0$,*

$$\Pr\left(0 < \min_{\lambda \in \Lambda(X;\theta):\lambda \neq \lambda^*(X;\theta,\tau_0)} (\lambda - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) \leq \delta\right) \leq C\delta.$$

First, by Assumption 6, (17) is $o_p(1)$ if the stochastic equicontinuity property holds: for all positive values $\delta_N = o(1)$,

$$\sup_{\|\tau - \tau_0\|_\infty \leq \delta_N} \left|\mathbb{G}_n\Big(\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\tau) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\})\lambda^\top I(Y)\Big)\right| = o_p(1).$$

To this end, note that by Assumption 3,

$$\Big|\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\tau) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\})\lambda^\top I(Y)\Big| \leq M \cdot 1\{\lambda^*(X;\theta,\tau) \neq \lambda^*(X;\theta,\tau_0)\},$$

where

$$\begin{aligned}
& 1\{\lambda^*(X;\theta,\tau) \neq \lambda^*(X;\theta,\tau_0)\} \\
= \; & 1\{0 < (\lambda^*(X;\theta,\tau) - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) < (\lambda^*(X;\theta,\tau) - \lambda^*(X;\theta,\tau_0))^\top (\tau_0(X) - \tau(X))\} \\
\leq \; & 1\Big\{0 < \min_{\lambda \in \Lambda(X;\theta):\lambda \neq \lambda^*(X;\theta,\tau_0)} (\lambda - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) \leq M\|\tau - \tau_0\|_\infty\Big\}
\end{aligned}$$

It follows that

$$\begin{aligned}
& E\Big[\sup_{\|\tau - \tau_0\|_\infty \leq \delta_N} \Big|\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\tau) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\})\lambda^\top I(Y)\Big|\Big] \\
\leq \; & \Pr\left(0 < \min_{\lambda \in \Lambda(X;\theta):\lambda \neq \lambda^*(X;\theta,\tau_0)} (\lambda - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) \leq \delta_N\right).
\end{aligned}$$

By Lemma 1 and Theorem 3 of Chen, Linton, and Van Keilegom (2003), (17) is $o_p(1)$. Second, for (18), observe that

$$\begin{aligned}
& E\Big[\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\hat{\tau}) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\})\lambda^\top I(Y)\Big|\hat{\tau}\Big] \\
= \; & E\Big[\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\hat{\tau}) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\})\lambda^\top \tau_0(X)\Big|\hat{\tau}\Big] \\
= \; & E[(\lambda^*(X;\theta,\hat{\tau}) - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) 1\{(\lambda^*(X;\theta,\hat{\tau}) - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) > 0\}|\hat{\tau}]
\end{aligned}$$

$$
\begin{aligned}
\leq\ & E\Big[(\lambda^*(X;\theta,\hat{\tau}) - \lambda^*(X;\theta,\tau_0))^\top(\tau_0(X) - \hat{\tau}(X)) \\
& \cdot 1\{0 < (\lambda^*(X;\theta,\hat{\tau}) - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) < (\lambda^*(X;\theta,\hat{\tau}) - \lambda^*(X;\theta,\tau_0))^\top(\tau_0(X) - \hat{\tau}(X))\}\Big|\hat{\tau}\Big] \\
\leq\ & M\|\hat{\tau} - \tau_0\|_\infty \Pr\Big(0 < \min_{\lambda\in\Lambda(X;\theta):\lambda\neq\lambda^*(X;\theta,\tau_0)} (\lambda - \lambda^*(X;\theta,\tau_0))^\top \tau_0(X) \leq M\|\hat{\tau} - \tau_0\|_\infty \Big|\hat{\tau}\Big) \\
\leq\ & CM^2\|\hat{\tau} - \tau_0\|_\infty^2,
\end{aligned}
$$

where the last inequality follows from Lemma 1. Then, by Assumption 6, (18) is $o_p(1)$. Now I can apply the central limit theorem to (16) to obtain the desired result. □

## Appendix B Monte Carlo Simulation

I consider the same data generating process as in Section 6 with $J = 2$. Fixing a counterfactual value $\underline{x} = (-0.5, 1)$ for $X_{it}$, I construct confidence intervals for the sharp bounds on the counterfactual choice probability $\Pr(Y_{it}(\underline{x}) = 1)$. I estimate observed conditional choice probabilities $\tau_0(x)$ from a logistic regression of $Y_{it}$ on $(X_{it}^{(1)}, X_{it}^{(2)}, \frac{1}{T}\sum_{t=1}^T X_{it}^{(1)}, \frac{1}{T}\sum_{t=1}^T X_{it}^{(2)})$. I normalize $\beta_0^{(1)}$ to one and estimate $\beta_0^{(2)}$ using the maximum score estimator. Then, I employ the model-based bootstrap procedure proposed by Cattaneo et al. (2020) to construct confidence intervals for $\beta_0^{(2)}$, where I plug in the true Hessian matrix. I choose $N = 5000$ and $T = 10$, with $S = 1000$ simulations and $B = 399$ bootstrap replications. I set the nominal level $\alpha = 0.05$ and use $\delta = 0.025$ for the construction of confidence intervals for $\beta_0^{(2)}$.

In Table 5, I report the coverage rates and average excess lengths of the confidence intervals for sharp bounds on $\Pr(Y_{it}(\underline{x}) = 1)$ as proposed in Proposition 1. For comparison, I also consider infeasible scenarios where the true values of $\tau_0(x)$ and/or $\beta_0^{(2)}$ are known. One can see that the coverage rates are above 95% in most cases. The estimation error of $\tau_0$ has minimal effects on both the coverage and the length of the confidence intervals. In contrast, accounting for the estimation error of $\beta_0^{(2)}$ leads to conservative coverage and longer confidence intervals.

| | Coverage | Avg. Excess Length |
|---|---|---|
| true $\tau_0$, true $\beta_0^{(2)}$ | 0.963 | 0.023 |
| estimated $\tau_0$, true $\beta_0^{(2)}$ | 0.967 | 0.023 |
| true $\tau_0$, CI for $\beta_0^{(2)}$ | 0.997 | 0.073 |
| estimated $\tau_0$, CI for $\beta_0^{(2)}$ | 0.997 | 0.073 |

Table 5: 95% CI for sharp bounds on $\Pr(Y_{it}(\underline{x}) = 1)$