BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ESSAYS ON THE ECONOMETRIC ANALYSIS OF COUNTERFACTUAL POLICIES AND INCOMPLETENESS

by

YAN LIU

B.A., Peking University, 2015
M.A., Kyoto University, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2025

Approved by

First Reader

Hiroaki Kaido, PhD
Associate Professor of Economics

Second Reader

Iván Fernández-Val, PhD
Professor of Economics

Third Reader

Jean-Jacques Forneron, PhD
Assistant Professor of Economics

"It is the brain, the little grey cells on which one must rely. The senses mislead. One must seek the truth within—not without."

Agatha Christie

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my main advisor, Hiroaki Kaido, for his unwavering guidance, support, and encouragement throughout my PhD studies. I have also learned a lot from collaborating with him as his teaching assistant, research assistant, and coauthor. It has been a privilege to grow under such an exceptional mentor, both academically and personally.

I am also very grateful to my other committee members, Iván Fernández-Val and Jean-Jacques Forneron, for insightful discussions and invaluable suggestions. I thank Pierre Perron for his continued interest in my work and for sharing his general philosophy about econometrics research. He generously provided me with a research assistant opportunity that made the final stage of my PhD studies much smoother. I thank Zhongjun Qu, Marc Rysman, and many other faculty members and external speakers whose names I have not explicitly mentioned for their helpful comments.

I am thankful to my advisor during my master's studies at Kyoto University, Yoshihiko Nishiyama, who enlightened me to pursue an academia career as an econometrician and has believed in me all these years. I also thank Toru Kitagawa and Atsushi Kajii, without whose support I would not have embarked on this journey.

I am fortunate to have an amazing PhD cohort, among whom are Qingyuan Chai, Danrong Chen, Seoyun Stella Hong, Lei Ma, Luwen Mai, Zhongyi Tang, Jiaqi Yang, Huihan Zhang, and Pengyue Zhu. We have created lots of precious memories together over the years. I also thank my fellow (and alumni) econometricians in the program, including Undral Byambadalai, Mingli Chen, Shuowen Chen, Matthew Hong, Haoran Pan, Yi Zhang, and Liang Zhong, for many lighthearted yet stimulating discussions.

I thank my dear friends from my years back in China and Japan, especially Chao Jin and Jiadi Qin, who have shared parallel journeys in academia with me. Last but not least, my thanks go to my parents for their unconditional love and support.

ESSAYS ON THE ECONOMETRIC ANALYSIS OF COUNTERFACTUAL POLICIES AND INCOMPLETENESS

YAN LIU

Boston University, Graduate School of Arts and Sciences, 2025

Major Professor: Hiroaki Kaido, PhD
Associate Professor of Economics

ABSTRACT

This dissertation consists of three essays in econometrics, which are related through two common themes: counterfactual policies and incompleteness. Chapters 1 and 2 are about predicting the potential impact of counterfactual policies, with an emphasis on controlling for endogeneity using panel data or instrumental variables. Chapters 1 and 3 deal with incompleteness that stems from either incomplete data or incomplete models, employing partial identification approaches.

Chapter 1 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. 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.

Chapter 2 studies the identification and estimation of individualized intervention policies in observational data settings characterized by endogenous treatment selection and the availability of instrumental variables. I introduce encouragement rules that manipulate an instrument. Incorporating the marginal treatment effects (MTE) as policy invariant structural parameters, I establish the identification of the social welfare criterion for the optimal encouragement rule. Focusing on binary encouragement rules, I propose to estimate the optimal policy via the Empirical Welfare Maximization (EWM) method and derive convergence rates of the regret (welfare loss). I consider extensions to accommodate multiple instruments and budget constraints. Using data from the Indonesian Family Life Survey, I apply the EWM encouragement rule to advise on the optimal tuition subsidy assignment. My framework offers interpretability regarding why a certain subpopulation is targeted.

Chapter 3 (joint with Hiroaki Kaido) expands the scope of likelihood-based model selection tests to a broad class of discrete choice models. A notable feature is that each of the competing models can make either a complete or incomplete prediction. We provide a novel cross-fitted likelihood-ratio statistic for such settings, which can be compared to a normal critical value. The proposed test does not require any information on how an outcome is chosen when multiple solutions are predicted. This allows the practitioner to compare, for example, a model that predicts a unique equilibrium to another model that allows for multiple equilibria. We examine the finite-sample properties of the test and provide guidance on the choice of tuning parameters through Monte Carlo experiments.

Contents

List of Tables

List of Figures

List of Abbreviations

a.e.	.....	almost everywhere
a.s.	.....	almost surely
avg.	.....	average
BEWM	.....	Budget-Constrained Empirical Welfare Maximization
CCP	.....	conditional choice probability
CDF	.....	cumulative distribution function
CI	.....	confidence interval
ct	.....	counterfactual
DR	.....	doubly robust
est.	.....	estimated
FE	.....	fixed effects
FEWM	.....	Feasible Empirical Welfare Maximization
i.i.d.	.....	independent and identically distributed
LES	.....	Linear Eligibility Score
ob	.....	observed
$\mathbb{R}$	.....	the real line
sec.	.....	second(s)
s.t.	.....	such that/subject to
Std. Dev.	.....	standard deviation
Supp	.....	support
TA	.....	Threshold Allocations
UK	.....	United States
US	.....	United Kindom
VC	.....	Vapnik–Chervonenkis

Chapter 1

Robust Counterfactual Analysis for Nonlinear Panel Data Models

1.1 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 chapter aims to fill this gap.

This chapter 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 et al., 2013a) 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 (including binary choice, ordered choice, and 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 when structural parameters need to be estimated. I propose a bootstrap-based procedure and provide simulation evidence on its performance.

As an empirical illustration, I apply my approach to US and UK 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 chapter 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 et al. (2016, 2023), Shi et al. (2018), Gao and Li (2024), Khan et al. (2021), Botosaru et al. (2023), Chesher et al. (2024), Gao and Wang (2024), Pakes and Porter (2024). It is well known that struc-

tural 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 chapter complements the literature on the identification of counterfactuals in discrete outcome models, including Manski (2007), Chiong et al. (2021), Gu et al. (2024), Tebaldi et al. (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 chapter adds to the literature on the identification of counterfactuals in nonlinear panel data models, including Hoderlein and White (2012), Chernozhukov et al. (2013a), Chernozhukov et al. (2019), Liu et al. (2024), Davezies et al. (2024), 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. (2013a) 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. (2024) 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. (2024) 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 the chapter is organized as follows. Section 1.2 outlines the setup and specifies the type of counterfactuals under consideration. Section 1.3 presents the sharp identified set for the distribution of the counterfactual outcome. Section 1.4 discusses the tractable implementation of the sharp identified set. Section 1.5 addresses estimation and inference. Section 1.6 gives numerical results for the sharp identified set. Section 1.7 contains empirical illustrations using data on female labor force participation and purchases of saltine crackers. Section 1.8 explores the extension to dynamic binary choice models. Section 1.9 concludes. Proofs and simulation results are collected in Appendix A.

1.2 Setup

This chapter considers panel data models of the form:

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

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 θ_0. Write $X_i = (X_{i1}, \ldots, X_{iT})$. Throughout, I assume that the data are i.i.d. across i. For the identification analysis in Sections 1.3, 1.4, and 1.8 and the numerical experiments in Section 1.6, I drop the i subscript to simplify the notation.

Example 1.1 (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 1.2 (Ordered Choice Model)**.** *Consider the model*

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

where $\beta_0 \in \mathbb{R}^{d_x}$ is a vector of unknown coefficients, and $\gamma_0 = (\gamma_0^1, \ldots, \gamma_0^J)$ are unknown thresholds satisfying $\gamma_0^j > \gamma_0^{j-1}$ and γ_0^1 is normalized to 0. Here $\theta_0 = (\beta_0, \gamma_0)$ and $\mathcal{Y} = \{0, 1, \ldots, J\}$. When $J = 1$, the model reduces to Example 1.1.

Example 1.3 (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 1.4 (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

$$Y_{it} = \arg\max_{j \in \mathcal{Y}} (X_{jit}^\top \beta_0 + U_{jit}),$$

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 1.1.

Assumption 1.1 (Time Homogeneity)**.** *$U_{it} \overset{d}{=} U_{i1}|X_i$ for all t.*

Assumption 1.1 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. (2013a) 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 $V_{it} \overset{d}{=} V_{i1}|X_i, A_i$ for all t, and A_i is a time-invariant individual effect. It is worth noting that Assumption 1.1 excludes lagged Y_{it} from X_{it} and focuses on static models. On the other hand, Assumption 1.1 allows U_{it} to be correlated with X_i and dependent over time. Moreover, it places no parametric distributional restriction on U_{it}.

Assumption 1.2. *θ_0 is known or point-identified.*

Assumption 1.2 is satisfied for a broad class of structural functions g under Assumption 1.1 and rich support conditions for U_{it} and X_i. In particular, it holds for all the examples mentioned above. For Example 1.1, Manski (1987) showed the identification of θ_0 up to scale. For Example 1.2, Botosaru et al. (2023) showed the identification of θ_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 1.3, Honoré and Kyriazidou (2000) showed the identification of θ_0. For

Example 1.4, point identification of θ_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 θ_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 1.4. However, it is not used in deriving sharp identified sets of counterfactuals in Section 1.3. In other words, results in Section 1.3 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 1.2 and 1.3, 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 1.4, 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 1.1. *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 1.2. *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 chapter.*

1.3 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 $\mathcal{S} \subseteq \text{Supp}(W|X = x)$, $F_{W|X=x}(\mathcal{S}) = \Pr(W \in \mathcal{S}|X = x)$. Write $Y = (Y_1, \ldots, Y_T)$.

Define the U-level set as

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

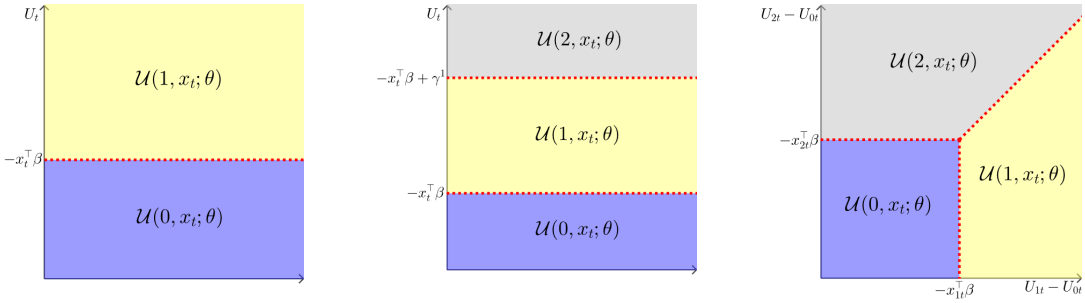
so that

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

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 1·1 contains stylized depictions of U-level sets in Examples 1.1, 1.2, and 1.4 with $J = 2$. For any measurable 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

[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).



Binary Choice Model Ordered Choice Model Multinomial Choice Model

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

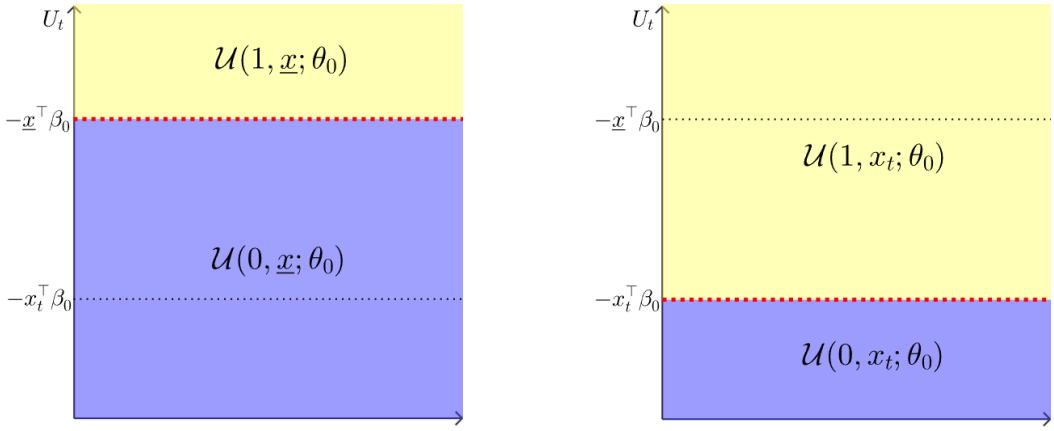
characterized as

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

where $\mathsf{F}(\mathcal{Y})$ denotes the collection of all measurable subsets of $\mathcal{Y}$. Therefore, to identify the distribution of $Y_t(\underline{x})$, it is necessary to identify θ_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 1.2, has been studied in the literature for a broad 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 1.1 using Figure 1·2.

As shown in Figure 1·2, 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 1.1 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 θ_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.



U-Level Sets under the Counterfactual U-Level Sets for the Observed Data

Figure 1·2: Discrepancy of U-Level Sets: Binary Choice Model

Given the impossibility of point identification, I provide the sharp identified set of the distribution of $Y_t(\underline{x})$ in Theorem 1.1. The proof is in Appendix A.1. 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 1.1 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 1.1. *Suppose that Assumptions 1.1 and 1.2 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*

$$\mathsf{F}^*_{Y_t(\underline{x})|X} = \{\mathcal{F}_{Y_t(\underline{x})|X} : \exists \mathcal{F}_{U_t|X} \in \mathsf{F}^*_{U_t|X} \\ s.t.\ \forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta_0))\ a.e.\ x \in \mathrm{Supp}(X)\}, \tag{1.1}$$

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

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

Remark 1.3. *Point identification of θ_0 (Assumption 1.2) is imposed to fix ideas and is stronger than necessary. The identified set defined in (1.1) is sharp for a given value of θ_0. When point identification of θ_0 fails, one can still take the union of (1.1) over the sharp identified set for θ_0 to obtain the sharp identified set for $\mathcal{F}_{Y_t(\underline{x})|X}$.*

1.4 Implementation

By Theorem 1.1, the most straightforward way to implement $\mathsf{F}^*_{Y_t(\underline{x})|X}$ is to search over the space of distributions supported on

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

for each $x \in \mathrm{Supp}(X)$. 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

[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):

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

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 a separable index restriction on g, with a focus on Examples 1.1-1.4. I start with a heuristic illustration in Example 1.1.

As shown in Figure 1·2, 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:

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

In this way, I generate identifying restrictions on $\mathcal{F}_{Y_t(\underline{x})|X}$ directly from $\mathcal{F}_{Y_t|X}$. Under Assumption 1.1, 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 1.2, a common structure in Examples 1.1-1.4 makes it easier to determine these set inclusion relationships. More formally, Examples 1.1-

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

1.4 satisfy an *index separability* condition in the sense that by partitioning $\theta = (\beta, \gamma)$,

$$(\mathcal{T}, \mathcal{T}') \in \mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma) \;\Rightarrow\; \mathcal{U}(\mathcal{T}, \underline{x}; \theta) \subseteq \mathcal{U}(\mathcal{T}', x_t; \theta) \tag{1.2}$$

for some collection $\mathbb{Y}(\underline{x}^\top \beta, x_t^\top \beta; \gamma)$ of pairs of subsets of $\mathcal{Y}$.[3] Here, β represents the index coefficients and γ collects the remaining parameters. In words, (1.2) 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 1.1-1.3 are encompassed by the following *monotone transformation model*.

Example 1.5 (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 γ_0. For Example 1.1, $h(v; \gamma) = 1\{v \geq 0\}$. For Example 1.2, $h(v; \gamma) = \sum_{j=1}^{J} 1\{v \geq \gamma^j\}$. For Example 1.3, $h(v; \gamma) = \max\{0, v\}$. 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

$$\mathcal{U}([y, \infty), x_t; \theta) = [-x_t^\top \beta + h^-(y; \gamma), \infty). \tag{1.3}$$

[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 θ. However, in this chapter, I focus on linear indices that are the most commonly used in practice.

Also define

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

One can predict the following set inclusion relationships:

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

Example 1.4 (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 γ is not present in this example, I omit it and define

$$\begin{aligned}\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) = \Big\{&(\mathcal{T},\mathcal{T}) : \mathcal{T}\subsetneq\{0,1,\ldots,J\}, \mathcal{T}\neq\emptyset, \\ &\min_{j\in\mathcal{T}}(x_{jt}-\underline{x}_j)^\top\beta \geq \max_{k\notin\mathcal{T}}(x_{kt}-\underline{x}_k)^\top\beta\Big\}. \qquad (1.4)\end{aligned}$$

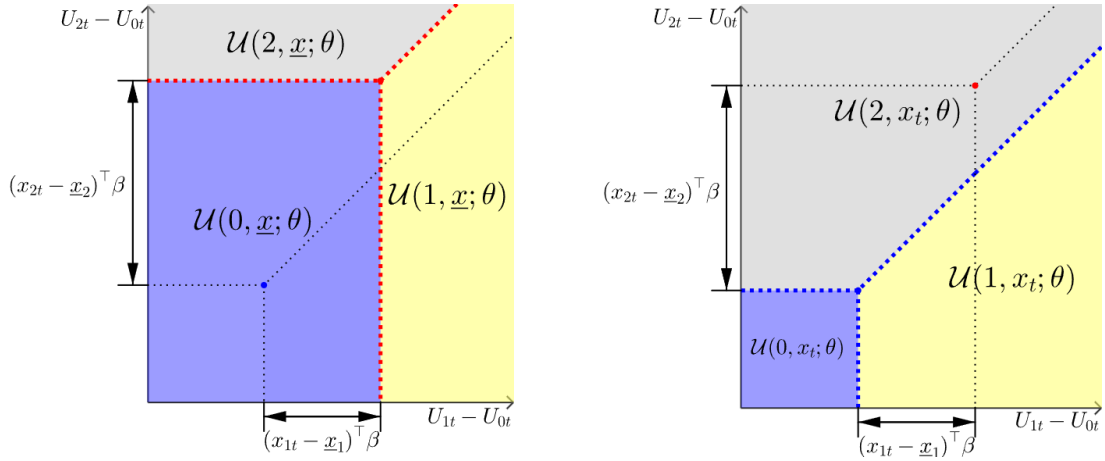
Intuitively, for any $\mathcal{T}$ satisfying the restrictions in (1.4), moving from $\underline{x}$ to x_t makes alternatives in $\mathcal{T}$ more likely to be chosen, regardless of the distribution of U_t. Hence, one can predict the following set inclusion relationships:

$$(\mathcal{T},\mathcal{T}')\in\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) \;\Rightarrow\; \mathcal{U}(\mathcal{T},\underline{x};\theta)\subseteq\mathcal{U}(\mathcal{T}',x_t;\theta). \qquad (1.5)$$

A proof of relation (1.5) is given in Appendix A.1. It is helpful to understand (1.5) 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, $\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta) = \{(\{2\},\{2\}),(\{2,1\},\{2,1\})\}$. As shown in Figure 1·3, there are two set inclusion relationships:

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

In general, to construct $\mathbb{Y}(\underline{x}^\top\beta, x_t^\top\beta)$, one can simply rank the $J+1$ index function



U-Level Sets under the Counterfactual U-Level Sets for the Observed Data

Figure 1·3: Set Inclusion Relationships of U-Level Sets: Multinomial Choice Model

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 1.5 and 1.4 in Theorems 1.2 and 1.3, respectively. The proofs are in Appendix A.1.

Theorem 1.2. *Suppose that Assumptions 1.1 and 1.2 hold. Let* g *be specified as in Example 1.5. Then,*

$$\mathsf{F}^*_{Y_t(\underline{x})|X} = \bigcap_{t'=1}^{T} \Big\{ \mathcal{F}_{Y_t(\underline{x})|X} : \forall (\mathcal{T}, \mathcal{T}') \in \mathbb{Y}_u(\underline{x}^\top \beta_0, x_{t'}^\top \beta_0; \gamma_0), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \leq F_{Y_{t'}|X=x}(\mathcal{T}'), \\ \forall (\mathcal{T}, \mathcal{T}') \in \mathbb{Y}_l(\underline{x}^\top \beta_0, x_{t'}^\top \beta_0; \gamma_0), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \geq F_{Y_{t'}X=x}(\mathcal{T}') \text{ a.e. } x \in \text{Supp}(X) \Big\}. \tag{1.6}$$

By Theorem 1.2, the sharp bounds on the counterfactual survival probability $F_{Y_t(\underline{x})|X=x}([y, \infty))$ are given by

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

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 1.3. *Suppose that Assumptions 1.1 and 1.2 hold. Let g be specified as in Example 1.4. Then,*

$$\mathsf{F}^*_{Y_t(\underline{x})|X} = \bigcap_{t'=1}^{T} \{\mathcal{F}_{Y_t(\underline{x})|X} : \forall (\mathcal{T}, \mathcal{T}') \in \mathbb{T}(\underline{x}^\top \beta_0, x_{t'}^\top \beta_0), \\ F_{Y_t(\underline{x})|X=x}(\mathcal{T}) \leq F_{Y_{t'}|X=x}(\mathcal{T}') \text{ a.e. } x \in \text{Supp}(X)\}. \tag{1.7}$$

A collection of choice sets similar to (1.4) 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 θ_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 θ_0.

1.5 Estimation and Inference

Notation Let $\|\cdot\|$ denote the Euclidean norm. Let $\rightsquigarrow$ and $\overset{p}{\rightsquigarrow}$ denote weak convergence and weak convergence in probability, respectively.

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

$$[\Psi_l(\theta_0), \Psi_u(\theta_0)] = \Big[E\Big[\max_{\lambda\in\Lambda_l(X;\theta_0)} \lambda^\top \tau_0(X)\Big], E\Big[\min_{\lambda\in\Lambda_u(X;\theta_0)} \lambda^\top \tau_0(X)\Big]\Big], \tag{1.8}$$

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 (1.8). 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 yth 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 tth place.

Example 1.2 (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 (1.8). To see this, note that by Theorem 1.2, the bounds are given by*

$$\Big[E\Big[\max_t \psi_t^l(X;\theta_0)\Big], E\Big[\min_t \psi_t^u(X;\theta_0)\Big]\Big],$$

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,

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

Example 1.4 (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 (1.8). To see this, note that by Theorem 1.3, 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*

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

where Δ^{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:

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

Define

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

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 τ_0.

Definition 1.1 (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 τ_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 ith observation.*

I impose the following assumptions.

Assumption 1.3. *For all θ, $\max_{\lambda \in \Lambda_l(x;\theta) \cup \Lambda_u(x;\theta)} \|\lambda\| \leq M$ for some $M > 0$ a.e. $x \in \text{Supp}(X)$.*

Assumption 1.4. *For all θ and τ, $\arg\max_{\lambda \in \Lambda_l(x;\theta)} \lambda^\top \tau(x)$ and $\arg\min_{\lambda \in \Lambda_u(x;\theta)} \lambda^\top \tau(x)$ are singletons a.e. $x \in \text{Supp}(X)$.*

Assumption 1.5. *The distribution of $\tau_0(X)$ is absolutely continuous with density bounded above.*

Assumption 1.6. *$\|\hat{\tau} - \tau_0\|_\infty = o_p(N^{-1/4})$, where $\|\tau\|_\infty = \sup_x \|\tau(x)\|$.*

Assumption 1.3 imposes boundedness on the objective function of the optimization problems and is satisfied in Examples 1.2 and 1.4. Assumption 1.4 requires the solution of the optimization problems to be unique. Assumption 1.5 is a sufficient condition for the margin condition (Lemma A.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 1.5 because it is low-level and compatible with the sufficient conditions for Assumption 1.2. Assumption 1.6 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 et al. (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

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

Given the first-step cross-fitted estimator $\hat{\tau}$ of τ_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 1.4 establishes the pointwise asymptotic normality of $\hat{\Psi}_l(\theta)$ and $\hat{\Psi}_u(\theta)$.

Theorem 1.4. *Suppose that Assumptions 1.3-1.6 hold. Then, for a given θ,*

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

where

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

With a first-step estimate $\hat{\theta}$ of θ_0, feasible estimators of $\Psi_l(\theta_0)$ and $\Psi_u(\theta_0)$ can be given by $\hat{\Psi}_l(\hat{\theta})$ and $\hat{\Psi}_u(\hat{\theta})$. However, their asymptotic distribution is complicated by the estimation error of $\hat{\theta}$. I give a heuristic explanation for $\hat{\Psi}_l(\hat{\theta})$ in Example 1.1. 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 1.4, $\hat{\Psi}_l(\hat{\theta}) - \Psi_l(\hat{\theta}) = O_p(N^{-1/2})$. Note that θ enters $\Psi_l(\theta)$ only through

Λ_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, $\text{sgn}((x_t - \underline{x})^\top \beta) \neq \text{sgn}((x_t - \underline{x})^\top \beta_0)$, 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. I focus on cases where $\hat{\theta}$ exhibits general cube root asymptotics and a bootstrap-based distributional approximation is available.

Assumption 1.7. *For some $q_N > 0$ with $r_N = (Nq_N)^{1/3} \to \infty$, a bootstrap version $\tilde{\theta}^*$ of $\hat{\theta}$, and a Gaussian process $Z(s)$, (i) $r_N(\hat{\theta} - \theta_0) \rightsquigarrow \arg\max_{s \in \mathbb{R}^{d_\theta}} Z(s)$, (ii) $r_N(\tilde{\theta}^* - \hat{\theta}) \overset{p}{\rightsquigarrow} \arg\max_{s \in \mathbb{R}^{d_\theta}} Z(s)$.*

Example 1.5. *For the panel maximum score estimator of θ_0, Cattaneo et al. (2020) provided primitive conditions under which Assumption 1.7 holds with $q_N = 1$ and showed that $\tilde{\theta}^*$ can be constructed by analytically modifying the criterion function.*

Example 1.4. *Khan et al. (2021) proposed a localized maximum score estimator of θ_0, for which Assumption 1.7(i) holds with $q_N = h_N^{(J-1)d_\theta}$ for $h_N > 0$ such that as $N \to \infty$, $h_N \to 0$, $Nh_N^{(J-1)d_\theta} / \log N \to \infty$ and $Nh_N^{(J-1)d_\theta + 3} \to 0$. However, it is unclear how to construct $\tilde{\theta}^*$ using bootstrap-based procedures.*

A natural conjecture is that under Assumption 1.7, the distribution of $r_N(\hat{\Psi}_l(\hat{\theta}) - \Psi_l(\theta_0))$ (resp. $r_N(\hat{\Psi}_u(\hat{\theta}) - \Psi_u(\theta_0))$) can be consistently estimated by that of $r_N(\hat{\Psi}_l(\tilde{\theta}^*) - \hat{\Psi}_l(\hat{\theta}))$ (resp. $r_N(\hat{\Psi}_u(\tilde{\theta}^*) - \hat{\Psi}_u(\hat{\theta}))$). Then, one can construct a $(1-\alpha)$-confidence interval for $[\Psi_l(\theta_0), \Psi_u(\theta_0)]$ as

$$\text{CI}_N(\alpha) = [\hat{\Psi}_l(\tilde{\theta}^*)_{(\alpha/2)}, \hat{\Psi}_u(\tilde{\theta}^*)_{(1-\alpha/2)}], \tag{1.9}$$

where $\hat{\Psi}_l(\tilde{\theta}^*)_{(\alpha)}$ (resp. $\hat{\Psi}_u(\tilde{\theta}^*)_{(\alpha)}$) denotes the α-quantile of the distribution of $\hat{\Psi}_l(\tilde{\theta}^*)$ (resp. $\hat{\Psi}_u(\tilde{\theta}^*)$). A preliminary Monte Carlo experiment conducted in Appendix A.2 shows that the confidence interval in (1.9) has correct coverage for sufficiently large

N, which suggests that this inference procedure may be asymptotically valid under suitable conditions. I leave the formal analysis for future work.

Remark 1.4. *The confidence interval in (1.9) is for the sharp bounds* $[\Psi_l(\theta_0), \Psi_u(\theta_0)]$ *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 chapter.*

Remark 1.5. *The confidence interval in (1.9) is two-sided. If one is only interested in the lower (resp. upper) bound on the counterfactual probability, it is straightforward to construct a one-sided confidence interval by using* $\hat{\Psi}_l(\tilde{\theta}^*)_{(\alpha)}$ *(resp.* $\hat{\Psi}_u(\tilde{\theta}^*)_{(1-\alpha)}$*) instead of* $\hat{\Psi}_l(\tilde{\theta}^*)_{(\alpha/2)}$ *(resp.* $\hat{\Psi}_u(\tilde{\theta}^*)_{(1-\alpha/2)}$*) and setting the other side to 1 (resp. 0).*

1.6 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 1.2, I consider the following data generating process:

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

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, $X_t^{(2)} \sim$ 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 the specific expressions for (1.8) in Example 1.2 with the expectation approximated by 5,000 random draws. Figure 1·4 shows the sharp bounds on $\Pr(Y_t(\underline{x}) \geq 1)$ 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.

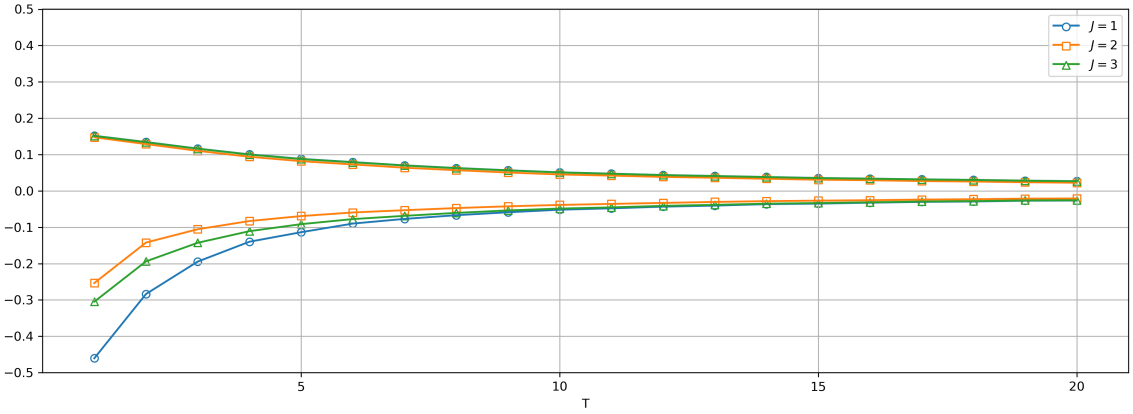


Figure 1·4: Re-centered Sharp Bounds on $\Pr(Y_t(\underline{x}) \geq 1)$ in Ordered Choice Models

For Example 1.4, 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 1.2, $X_{jt}^{(1)} \sim N(0, 0.5)\ \forall j$ 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, $X_{jt}^{(2)} \sim$ 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 the specific expressions for (1.8) in Example 1.4 with the expectation approximated by 5,000 random draws. Figure 1·5 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 1·4. Unlike in Figure 1·4, the bounds become wider when J increases. A plausible explanation is that unlike Example 1.2, here a larger J leads to higher-dimensional unobserved heterogeneity, whose distribution may require more data to learn about.

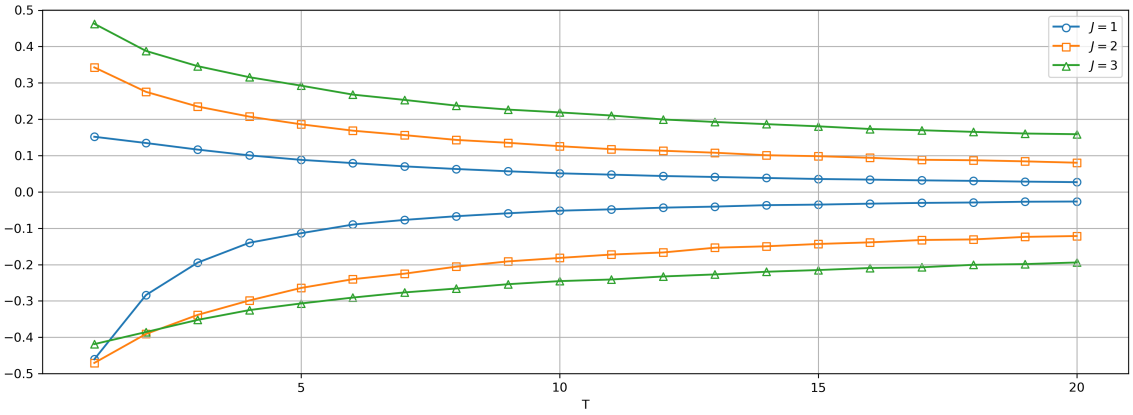


Figure 1·5: Re-centered Sharp Bounds on $\Pr(Y_t(\underline{x}) = 1)$ in Multinomial Choice Models

1.7 Empirical Applications

1.7.1 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 Carro (2007), 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_{it} is the labor force participation indicator, and X_{it} 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 1.1 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 1.1.

Table 1.1: 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 β_0. I estimate β_0 using the maximum-score-type objective function:

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

Table 1.2 reports the point estimates of β_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.

Consider the counterfactual scenario where log husband's income and age are set

Table 1.2: Estimated β_0

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

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 1.5 and plot them in Figure 1·6. To do this, I plug in the maximum score estimates of β_0 in Table 1.2 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%

[4]Note that under Assumption 1.1, 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 $\tau_0(x)$.

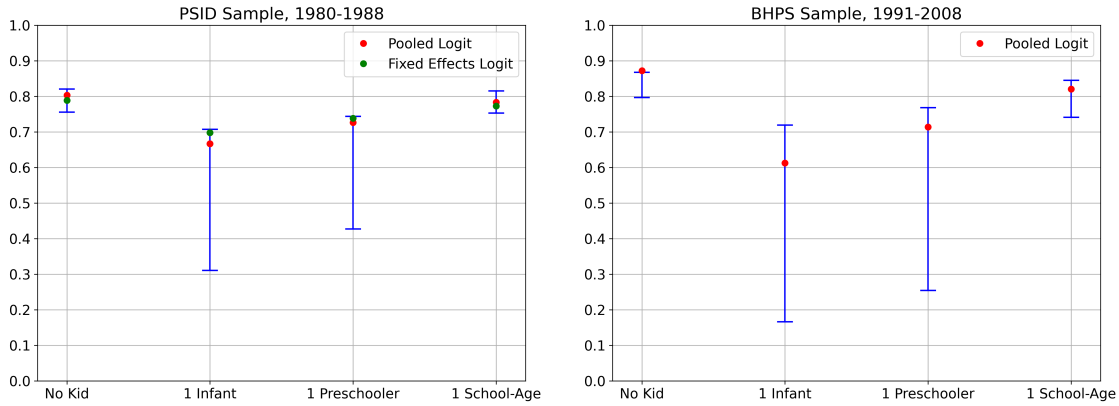


Figure 1·6: Counterfactual Probabilities of Labor Force Participation

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 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 β_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 1.2 under the columns "Pooled Logit" and "FE Logit". I plot predictions of counterfactual labor force participation rates from these two parametric specifications in Figure 1·6.[5] One can see that some parametric predictions lie close

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

to the upper bounds, suggesting that they may be overly optimistic.

1.7.2 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 is from Paap and Franses (2000) and was also analyzed in Khan et al. (2021). 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 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 1.3 reports the descriptive statistics for each brand.

Table 1.3: 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 1.1 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

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

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 α_j are parameters to be estimated.[6] Table 1.4 reports the point estimates of coefficients.[7]

Table 1.4: 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

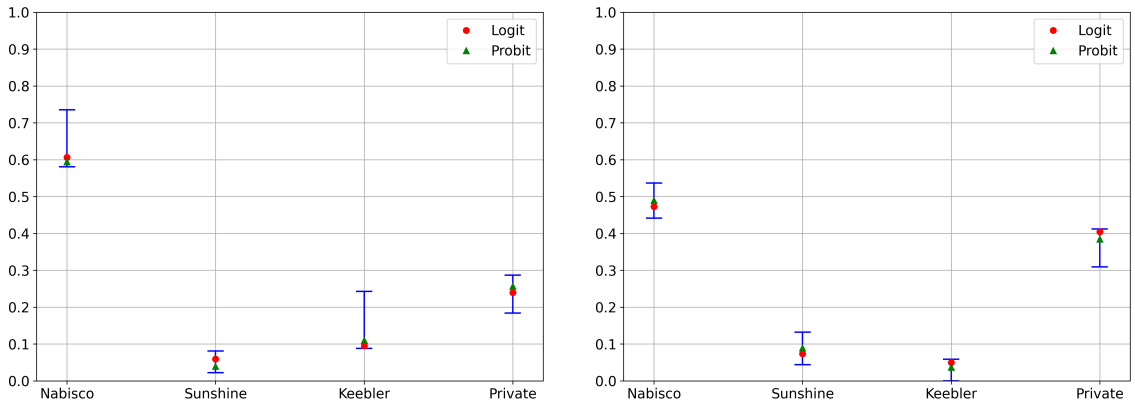
I consider the counterfactual choice probabilities under two counterfactual values $\underline{x}$ and $\overline{x}$ for explanatory variables. The price vector for $\underline{x}$ is $\underline{p} = (1.09, 1.05, 1.05, 0.78)$ and the price vector for $\overline{x}$ is $\overline{p} = (1.09, 0.89, 1.21, 0.59)$. 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 1.5. To do this, I plug in the semiparametric estimates of $(\beta_0^{(1)}, \beta_0^{(2)})$ in Table 1.4 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 1·7 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

[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 $X_{ijt}^{(1)}$.



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

Figure 1·7: Counterfactual Choice Probabilities

market share change for brand 2 (Sunshine) is ambiguous. For comparison, I also plot the predictions from pooled multinomial logit and probit models in Figure 1·7. 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).

1.8 Extension: Dynamic Binary Choice Models

Although the main framework of this chapter 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 outcome 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 1.1, which is termed *partial stationarity* in Gao and Wang (2024) because the conditioning set only contains part of the explanatory variables. Identification of θ_0 under Assumption 1.1 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 $y \in \{0, 1\}$,

$$U_t \in \mathcal{U}(y, Y_{t-1}, X_t; \theta_0) \text{ and } \mathcal{U}(y, Y_{t-1}, X_t; \theta_0) \subseteq \mathcal{U}(y, \underline{y}, \underline{x}; \theta_0) \;\Rightarrow\; U_t \in \mathcal{U}(y, \underline{y}, \underline{x}; \theta_0). \tag{1.10}$$

Note that

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

Taking the conditional expectation of (1.10) 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

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

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

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 1.1 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.

1.9 Conclusion

This chapter establishes the sharp identified set of the counterfactual outcome distribution 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.

Chapter 2

Policy Learning under Endogeneity Using Instrumental Variables

2.1 Introduction

An important goal of program evaluation is to inform policy on the assignment of individuals to treatment based on observable characteristics in the presence of treatment effect heterogeneity. Toward this goal, this chapter considers the statistical decision problem of learning an individualized intervention policy that maximizes social welfare. By leveraging an instrumental variable (IV), we address two issues with policy learning using data from observational studies or randomized experiments with noncompliance, which can be viewed as two sides of the same coin. First, when identifying causal effects to form the social welfare criterion, we allow for endogeneity (i.e., unobserved confoundedness). Second, when designing policies, we allow for imperfect enforcement. The remedy is a broader definition of policy interventions that encompasses not only directly altering treatment status but also inducing flows into (and possibly out of) treatment by manipulating the instrument. We refer to the latter as the *encouragement rule*, a term borrowed from the epidemiology literature.

The first of the above issues concerns the identification of the social welfare criterion to evaluate a policy. For example, the policymaker may wish to select a binary treatment rule that assigns enrollment in upper secondary school to maximize average adult wages. Each individual is assigned to a treatment based on observable

characteristics that affect wages (e.g., parental education, rural/urban residence). The crucial assumption to point-identify the social welfare criterion, defined as the average counterfactual outcome, is *unconfoundedness*, which states that the treatment is independent of potential outcomes conditional on observable characteristics. However, this assumption breaks down when selection is based on the unobserved components, e.g., ability or motivation, of heterogeneous responses to treatment. To deal with self-selection, we exploit an instrumental variable informative of marginal treatment selection behavior. This enables us to incorporate the concept of marginal treatment effects (MTE) (Björklund and Moffitt, 1987) as a building block of the social welfare criterion. In this sense, we bridge the literatures on the MTE and on statistical decision rules of policy interventions.

The second issue concerns the definition of policies. We note a fundamental asymmetry in the existing literature: the policymaker acknowledges endogenous treatment selection but designs treatment assignment rules as if they could be fully enforced. We address this asymmetry by expanding the universe of policies to include manipulations of the instrument that affect social welfare through the treatment, which we call encouragement rules. For instance, it is highly costly or impossible to force people to (or not to) attend upper secondary school. A more realistic scenario is to provide a scholarship or a tuition subsidy, whereas the tuition fee is commonly used as an instrument for school attendance.

We apply the social welfare criterion of an encouragement rule, identified via the MTE function, to one popular class of statistical decision rules: Empirical Welfare Maximization (EWM) rules (see Hirano and Porter, 2020, Section 2.3). We analyze the properties of the EWM method in terms of regret (welfare loss relative to an oracle optimal rule). To keep the analysis tractable, we focus on settings in which the policymaker makes binary decisions between two *a priori* chosen functions for

manipulation and constrains the class of feasible allocations.[1] Such settings are also of practical interest when the policymaker wants to avoid complicated rules or satisfy legal, ethical, or political considerations. In addition, we propose an alternative social welfare representation allowing for heterogeneity in treatment choice behavior when there are multiple instruments. As another practically relevant extension, we demonstrate how to incorporate budget constraints, which inherently involve unknown costs because of imperfect take-up.

We apply the EWM encouragement rule to an empirical dataset from the third wave of the Indonesian Family Life Survey (IFLS). We aim to provide advice on how upper secondary schooling can be encouraged to maximize average adult wages by manipulating the tuition fee. We find that the optimal policy without budget constraints provides tuition subsidy eligibility to individuals who face relatively high tuition fees and live relatively close to the nearest secondary school. The MTE structure underlying the social welfare criterion allows us to understand how the optimal policy is driven by heterogeneity in treatment take-up and treatment effects.

Related Literature: This chapter is related to four strands of literature, of which we provide a non-exhaustive overview below.

First, the research question is closely related to the literature on statistical treatment rules in econometrics, including Manski (2004), Hirano and Porter (2009), Bhattacharya and Dupas (2012), Kitagawa and Tetenov (2018b), Athey and Wager (2021), Mbakop and Tabord-Meehan (2021), Sun (2024). See also Hirano and Porter (2020) for a recent review. Despite the breadth of the literature, only a few econometric works look into policy choice with observational data when the unconfoundedness assumption does not hold. Kasy (2016) and Byambadalai (2022) focus on cases of partial identification and welfare ranking of policies rather than optimal policy choices.

[1] In principle, it is possible to generalize the regret analysis to multi-action settings. This would necessitate different tools, such as different complexity measures of the policy class, and is beyond the scope of this chapter.

Athey and Wager (2021) assume homogeneous treatment effects so that the conditional average treatment effect (CATE) on compliers can be extrapolated to those on the entire population. Sasaki and Ura (2024) identify the social welfare criterion via the MTE and demonstrate an application to the EWM framework. Nonetheless, these works implicitly assume the complete enforcement of treatment rules, whereas we consider more realistic policy tools.

The only exception we are aware of is Chen and Xie (2022), who use the MTE framework to study the personalized subsidy rule. Their work is complementary to ours in emphasizing different aspects of policy learning. They focus on the oracle optimal policy and its welfare properties without restricting the policy class. Notably, their characterization of optimal subsidy rules requires monotonicity of the propensity score and the MTE, which can be restrictive in practice. For example, in the context of the effects of family size on child outcomes, the MTE estimates in Brinch et al. (2017) show a U shape. In contrast, we do not have these requirements. Instead, we restrict the complexity of the policy class and analyze the convergence rate of the regret bound of the estimated policy.

Second, in epidemiology and biostatistics, there has been increasing interest in individualized treatment rules (ITR). Cui and Tchetgen Tchetgen (2021) and Qiu et al. (2021) allow for treatment endogeneity. They achieve point identification by leveraging the "no common effect modifier" assumption outlined in Wang and Tchetgen Tchetgen (2018), which largely restricts the heterogeneity of compliance behavior. Pu and Zhang (2021) introduce the notion of "IV-optimality" to estimate the optimal treatment regime based on partial identification of the CATE. Our approach crucially differs from these works on ITR under endogeneity in that we account for imperfect enforcement as a consequence of endogeneity. One exception is individualized encouragement rules that manipulate a binary instrument, considered in Qiu et al. (2021).

Our framework nests theirs.

Third, if one takes an *intention-to-treat* perspective, then optimization of the manipulation of a continuous instrument can also be studied under a framework of policy learning with continuous treatments, assuming unconfoundedness. Kallus and Zhou (2018) generalize the inverse propensity weighting method to the continuous treatment setting and propose a kernelized off-policy evaluator. Bertsimas and McCord (2018) focus on regression-based methods and propose predicted cost minimization penalized by variance and bias. Athey and Wager (2021) study binary decisions on whether or not to offer each individual an infinitesimal nudge to the preexisting treatment level. In contrast, our approach treats the variable intervened upon as an instrument and imposes the exclusion restriction that the instrument only affects the outcome through a binary treatment. The MTE structure, justified under the exclusion restriction, allows the policymaker to understand why the optimal policy targets a certain subpopulation and to rigorously perform extrapolation.

Lastly, our representation of the social welfare criterion can be viewed as a variation of policy relevant treatment effects (PRTE), adding to the class of policy parameters that can be written as weighted averages of the MTE. Hence, this chapter complements the literature on PRTE, including Carneiro et al. (2010, 2011), Mogstad et al. (2018), and Sasaki and Ura (2021) among many others.

Organization: The rest of the chapter is organized as follows. Section 2.2 sets up the model, introduces the encouragement rule, and derives a representation of the social welfare criterion via the MTE. Section 2.3 applies the social welfare representation to the EWM method and analyzes the regret properties. Section 2.4 discusses extensions incorporating multiple instruments and budget constraints. Section 2.5 presents an empirical application. Section 2.6 concludes. Appendix B.1 contains proofs of the main results. Appendix B.2 verifies the assumption of point identifica-

tion of the social welfare criterion under various specifications. Additional discussions and results are collected in the remaining sections of Appendix B.

2.2 Encouragement Rules with An Instrumental Variable

2.2.1 Setup

We consider the canonical program evaluation problem with a binary treatment $D \in \{0, 1\}$ and a scalar, real-valued outcome $Y \in \mathcal{Y} \subset \mathbb{R}$. Outcome production is modeled through the potential outcomes framework (Rubin, 1974):

$$Y = Y(1)D + Y(0)(1 - D),$$

where $(Y(0), Y(1))$ are the potential outcomes under no treatment and under treatment. Let $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote a vector of pretreatment covariates. For instance, in the analysis of returns to upper secondary schooling, D is an indicator for enrollment in upper secondary school, Y is the log wage, and X includes observable characteristics that affect wages (e.g., parental education, rural/urban residence). By adopting the potential outcomes model, we implicitly follow the conventional practice of imposing the Stable Unit Treatment Value Assumption (SUTVA), namely that there are no spillover or general equilibrium effects.

A (non-randomized) *treatment assignment rule* is defined as a mapping $\pi : \mathcal{X} \to \{0, 1\}$. The policymaker's objective function is the utilitarian (additive) welfare criterion defined by the average counterfactual outcome:

$$W(\pi) = E[Y(\pi(X))] = E[Y(1) \cdot \pi(X) + Y(0) \cdot (1 - \pi(X))].$$

Define the *conditional average treatment response* functions as $m_d(x) = E[Y(d)|X =$

$x], d = 0, 1$. By the law of iterated expectations,

$$W(\pi) = E[m_1(X) \cdot \pi(X) + m_0(X) \cdot (1 - \pi(X))].$$

Under the unconfoundedness assumption that D is independent of $(Y(0), Y(1))$ conditional on X, $m_d(x)$ is identified by $E[Y|D = d, X = x]$ for $d = 0, 1$. However, the unconfoundedness assumption is violated if, for example, people self-select into upper secondary education based on unmeasured benefits and costs driven by ability and motivation, both of which also affect wages. As a result, $m_d(x) \neq E[Y|D = d, X = x]$ for $d = 0, 1$ in general, and thus the social welfare criterion is not identified by the moments of observables. In this case, it is helpful to assume that there exists an instrument (i.e., an excluded variable) $Z \in \mathcal{Z} \subset \mathbb{R}$ that affects the treatment but not the outcome, e.g., a tuition subsidy. For each $z \in \mathcal{Z}$, denote the potential treatment status if the instrument were set to z by $D(z)$. The observed treatment is given by $D = D(Z)$. We explicitly model the selection into treatment via an additively separable latent index model:

$$D(z) = 1\{\tilde{\nu}(X, z) - \tilde{U} \geq 0\}, \tag{2.1}$$

where $\tilde{\nu}$ is an unknown function, and $\tilde{U}$ represents unobservable factors that affect treatment choice. Let "$\perp$" denote (conditional) statistical independence. We adopt the following assumptions from the MTE literature (Heckman and Vytlacil, 2005; Mogstad et al., 2018):

Assumption 2.1 (IV Restrictions and Continuous Distribution)**.** *(i)* $\tilde{U} \perp Z|X$. *(ii)* $E[Y(d)|X, Z, \tilde{U}] = E[Y(d)|X, \tilde{U}]$ *and* $E[|Y(d)|] < \infty$ *for* $d \in \{0, 1\}$. *(iii)* $\tilde{U}$ *is continuously distributed conditional on* X.

Assumptions 2.1(i) and (ii) impose exogeneity and an exclusion restriction on Z but allow for arbitrary dependence between $(Y(0), Y(1))$ and $\tilde{U}$, even conditional

on X. Vytlacil (2002) shows that, under Assumption 2.1(i), the existence of an additively separable selection equation as in (2.1) is equivalent to the monotonicity assumption used for the local average treatment effects (LATE) model of Imbens and Angrist (1994). The LATE monotonicity assumption restricts choice behavior in the sense that, conditional on X, an exogenous shift in Z either weakly encourages or discourages every individual to choose $D = 1$. Nonetheless, we maintain the selection equation (2.1) because it allows us to express the welfare contrast under counterfactual policies as a function of identifiable and interpretable objects. Under Assumptions 2.1(i) and (iii), we can reparameterize the model as

$$D(z) = 1\{\nu(X, z) - U \geq 0\} \quad \text{with} \quad U|X, Z \sim \text{Unif}[0, 1], \tag{2.2}$$

where $U \equiv F_{\tilde{U}|X}(\tilde{U}|X)$, $\nu(x, z) \equiv F_{\tilde{U}|X}(\tilde{\nu}(x, z)|x)$, and $\text{Unif}[a, b]$ denotes the uniform distribution over $[a, b]$. As a consequence,

$$p(x, z) \equiv \Pr(D = 1|X = x, Z = z) = F_{U|X,Z}(\nu(x, z)) = \nu(x, z),$$

where $p(x, z)$ is the propensity score.

Endogenous treatment selection also challenges the plausibility of fully mandating the treatment. We expand the universe of policies to include manipulations of the instrument that affect social welfare through the treatment. An *encouragement rule* is a mapping $\boldsymbol{\alpha} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ that manipulates the instrument for an individual with $(X, Z) = (x, z)$ from the initial value z to a new level $\boldsymbol{\alpha}(x, z)$. For example, when Z is the tuition fee, $\boldsymbol{\alpha}(x, z) = (z - \alpha(x)) \cdot 1\{z \geq \alpha(x)\}$ with $\alpha : \mathcal{X} \to \mathbb{R}_+$ describes a tuition subsidy rule that subsidizes an individual with $X = x$ up to $\alpha(x)$. Our representation of the social welfare criterion in Section 2.2.2 covers the most general setting without restrictions on $\boldsymbol{\alpha}$. When we study the regret bounds for statistical decision rules in Section 2.3, we take a stand on the complexity of feasible encouragement rules.

In particular, we focus on a case in which the policymaker makes a binary decision between two *a priori* chosen functions for manipulation. The binary formulation of encouragement rules nests treatment assignment rules (Kitagawa and Tetenov, 2018b; Sasaki and Ura, 2024) as a special case.

2.2.2 Representation of the Social Welfare Criterion via the MTE

The outcome that would be observed under encouragement rule $\boldsymbol{\alpha}$ is

$$Y(D(\boldsymbol{\alpha}(X,Z))) = Y(1)\cdot D(\boldsymbol{\alpha}(X,Z)) + Y(0)\cdot(1 - D(\boldsymbol{\alpha}(X,Z))).$$

We define the social welfare criterion as the average counterfactual outcome: $W(\boldsymbol{\alpha}) \equiv E[Y(D(\boldsymbol{\alpha}(X,Z)))]$. Theorem 2.1 shows that $W(\boldsymbol{\alpha})$ can be expressed as a function of the average treatment effect conditional on selection unobservables U and observable characteristics X, which is the definition of the MTE:

$$\mathrm{MTE}(u,x) \equiv E[Y(1) - Y(0) | U = u, X = x].$$

A proof is provided in Appendix B.1. The concept of MTE was introduced by Björklund and Moffitt (1987) and extended by Heckman and Vytlacil (1999, 2001b, 2005). In contrast to the intention-to-treat approach, the representation in Theorem 2.1 has a straightforward interpretation: among the individuals with $(X,Z) = (x,z)$, those for whom u is between $p(x,\boldsymbol{\alpha}(x,z))$ and $p(x,z)$ get either encouraged or discouraged to take up the treatment, and their contribution to the welfare contrast $W(\boldsymbol{\alpha}) - E[Y]$ is $\mathrm{MTE}(u,x)$ if encouraged and $-\mathrm{MTE}(u,x)$ if discouraged.

Theorem 2.1. *Under Assumption 2.1, the social welfare criterion for a given encouragement rule $\boldsymbol{\alpha}$ is given by*

$$W(\boldsymbol{\alpha}) = E[Y] + E\Big[\int_0^1 \mathrm{MTE}(u,X)\cdot(1\{p(X,\boldsymbol{\alpha}(X,Z)) \geq u\} - 1\{p(X,Z)\geq u\})\,du\Big].$$

A similar representation result appears in Chen and Xie (2022, Proposition 1), where they exclude Z from the targeting variables. This exclusion narrows their focus to policies that induce a degenerate distribution of Z conditional on $X = x$. In contrast, our framework also accommodates policies that shift the conditional distribution of Z through deterministic transformations, e.g., $\boldsymbol{\alpha}(x, z) = z + \alpha(x)$ with $\alpha : \mathcal{X} \to \mathbb{R}$.

Theorem 2.1 implies that the point-identification of $W(\boldsymbol{\alpha})$ is guaranteed by the point-identification of the propensity score and the MTE over necessary domains. We formalize this insight in the following assumption.

Assumption 2.2 (Point-Identification of $W(\boldsymbol{\alpha})$)**.** *(i) $p(x, z)$ is point-identified over* $\text{Supp}(X, \boldsymbol{\alpha}(X, Z))$*. (ii) For every $x \in \mathcal{X}$,* $\text{MTE}(\cdot, x)$ *is point-identified over* $[\min \mathcal{P}_{\boldsymbol{\alpha}}(x), \max \mathcal{P}_{\boldsymbol{\alpha}}(x)]$*, where $\mathcal{P}_{\boldsymbol{\alpha}}(x)$ denotes the support of $p(X, \boldsymbol{\alpha}(X, Z))$ conditional on $X = x$.*

We give three examples of sufficient conditions for Assumption 2.2. The verification of Assumption 2.2 in these examples is relegated to Appendix B.2. Typically, we need additional structural restrictions to compensate for the relaxation of the support condition.

Example 2.1 (Nonparametric Identification)**.** *Assume that (E2.1-1)* $\text{Supp}(X, \boldsymbol{\alpha}(X, Z)) \subset \text{Supp}(X, Z)$*; (E2.1-2) the conditional distribution of $p(X, Z)$ given X is absolutely continuous with respect to the Lebesgue measure. Then, Assumption 2.2 is satisfied. In particular,* $\text{MTE}(u, x)$ *can be point-identified by the method of local instrumental variables (LIV) (Heckman and Vytlacil, 1999, 2001a): for any $(x, u) \in$* $\text{Supp}(X, p(X, Z))$*,*

$$\text{MTE}(u, x) = \frac{\partial}{\partial u} E[Y | X = x, p(X, Z) = u].$$

As a result, $W(\boldsymbol{\alpha})$ is point-identified as

$$W(\boldsymbol{\alpha}) = E[Y] + E[g_Y(X, p(X, \boldsymbol{\alpha}(X, Z))) - g_Y(X, p(X, Z))],$$

where $g_Y(x, u) \equiv E[Y | X = x, p(X, Z) = u]$.

Example 2.2 (Semiparametric Identification). *Assume that (E2.2-1)* $\text{Supp}(\boldsymbol{\alpha}(X,Z)) \subset \mathcal{Z}$*; (E2.2-2) the distribution of* $p(X,Z)$ *is absolutely continuous with respect to the Lebesgue measure; (E2.2-3) the propensity score is modeled as* $p(x,z) = x^\top\gamma + \theta(z)$*, where* γ *is an unknown parameter and* θ *is an unknown function; (E2.2-4) the potential outcomes are modeled as* $Y(d) = X^\top\beta_d + V_d$ *for* $d = 0,1$*, where* β_1 *and* β_0 *are unknown parameters; (E2.2-5)* $E[V_d|X,Z,U] = E[V_d|U]$ *for* $d = 0,1$.[2] *Let* $V = DV_1 + (1-D)V_0$*. Then, Assumption 2.2 is satisfied, and* $W(\boldsymbol{\alpha})$ *is point-identified as*

$$\begin{aligned} W(\boldsymbol{\alpha}) = E[Y] &+ E[(p(X,\boldsymbol{\alpha}(X,Z)) - p(X,Z))X]^\top(\beta_1 - \beta_0) \\ &+ E[g_V(p(X,\boldsymbol{\alpha}(X,Z))) - g_V(p(X,Z))], \end{aligned}$$

where $g_V(x,u) \equiv E[V|p(X,Z) = u]$*. When (E2.2-1) is violated, Assumption 2.2 can still be satisfied by imposing a parametric model for the propensity score and a semiparametric partially linear model for the MTE, which is what we implement in the empirical application.*

Example 2.3 (Parametric Identification). *Assume that (E2.3-1) the propensity score is modeled as* $p(x,z) = \mu(x,z)^\top\gamma$*, where* μ *is a known vector function and* γ *is an unknown parameter;*[3] *(E2.3-2) the conditional mean of* Y *given* $(X, p(X,Z))$ *is modeled as* $E[Y|X = x, p(X,Z) = u] = ux^\top\beta_1 + (1-u)x^\top\beta_0 + \sum_{j=2}^J \eta_j u^j$*. Then, Assumption 2.2 is satisfied, and* $W(\boldsymbol{\alpha})$ *is point-identified as*

$$\begin{aligned} W(\boldsymbol{\alpha}) = E[Y] &+ E[(p(X,\boldsymbol{\alpha}(X,Z)) - p(X,Z))X]^\top(\beta_1 - \beta_0) \\ &+ \sum_{j=2}^J \eta_j E[p(X,\boldsymbol{\alpha}(X,Z))^j - p(X,Z)^j]. \end{aligned}$$

Polynomial MTE models are often used in empirical studies; see, e.g., Brinch et al. (2017), Cornelissen et al. (2018).

It is worth noting that $W(\boldsymbol{\alpha})$ has a natural connection to the concept of policy relevant treatment effects (PRTE) (Heckman and Vytlacil, 2005). For a general class

[2] The additive separability between observed and unobserved components and linear-in-parameters forms in (E2.2-4) are commonly assumed for potential outcomes in applied work estimating the MTE; see, e.g., Carneiro and Lee (2009), Carneiro et al. (2010, 2011). These works also invoke full independence $(U, V_0, V_1) \perp (Z, X)$, which is stronger than the mean independence of (V_0, V_1) from (Z, X) in (E2.2-5).

[3] Alternatively, one may adopt a logit or probit model to respect the $[0,1]$ boundary.

of policies that affect the propensity score, the PRTE is defined as the mean effect of going from a baseline policy to an alternative policy per net person shifted (assuming that $E[D|\text{alternative policy}] - E[D|\text{baseline policy}] \neq 0$):

$$\frac{E[Y|\text{alternative policy}] - E[Y|\text{baseline policy}]}{E[D|\text{alternative policy}] - E[D|\text{baseline policy}]}.$$

Corollary 2.1 gives an alternative representation of $W(\boldsymbol{\alpha})$ in terms of a suitably defined PRTE.

Corollary 2.1. *Under Assumption 2.1, when* $E[p(X, \boldsymbol{\alpha}(X, Z))] - E[p(X, Z)] \neq 0$, *the social welfare criterion for a given encouragement rule* $\boldsymbol{\alpha}$ *is given by*

$$W(\boldsymbol{\alpha}) = E[Y] + (E[p(X, \boldsymbol{\alpha}(X, Z))] - E[p(X, Z)]) \cdot \text{PRTE}(\boldsymbol{\alpha}),$$

where

$$\text{PRTE}(\boldsymbol{\alpha}) = E\Big[\int_0^1 \text{MTE}(u, X) \cdot \omega(u, X, Z; \boldsymbol{\alpha})\, du\Big]$$

with the weight defined as

$$\omega(u, x, z; \boldsymbol{\alpha}) \equiv \frac{1\{p(x, \boldsymbol{\alpha}(x, z)) \geq u\} - 1\{p(x, z) \geq u\}}{E[p(X, \boldsymbol{\alpha}(X, Z))] - E[p(X, Z)]}.$$

Corollary 2.1 unfolds the two forces driving the optimal policy based on $W(\boldsymbol{\alpha})$: the average change in treatment take-up, $E[p(X, \boldsymbol{\alpha}(X, Z))] - E[p(X, Z)]$, and the average treatment effect among those induced to switch treatment status, $\text{PRTE}(\boldsymbol{\alpha})$, when going from the status quo to encouragement rule $\boldsymbol{\alpha}$. Moreover, $\text{PRTE}(\boldsymbol{\alpha})$ can be expressed as a weighted average of the MTE with weights determined by both observed and unobserved heterogeneity in treatment take-up.

2.2.3 Binary Encouragement Rules

When we study the regret bounds for statistical decision rules in Section 2.3, we focus on settings in which the policymaker *a priori* chooses two functions $\alpha_0, \alpha_1 : \mathcal{Z} \to \mathbb{R}$

and decides between manipulating the instrument according to α_0 or α_1. A *binary encouragement rule*, indexed by a mapping $\pi : \mathcal{X} \times \mathcal{Z} \to \{0, 1\}$, manipulates the instrument for an individual with $(X, Z) = (x, z)$ to

$$\boldsymbol{\alpha}^{\pi}(x, z) = \pi(x, z) \cdot \alpha_1(z) + (1 - \pi(x, z)) \cdot \alpha_0(z).$$

Corollary 2.2 specializes Theorem 2.1 to the binary setting.

Corollary 2.2. *Under Assumption 2.1, we have*

$$W(\boldsymbol{\alpha}^{\pi}) = E[Y(D(\alpha_0(Z)))] + E\Big[\pi(X, Z) \cdot \int_0^1 \text{MTE}(u, X) \cdot (1\{p(X, \alpha_1(Z)) \geq u\} - 1\{p(X, \alpha_0(Z)) \geq u\})\, du\Big].$$

Henceforth, with a slight abuse of notation, we write $\int_{p(X,\alpha_0(Z))}^{p(X,\alpha_1(Z))} \text{MTE}(u, X)\, du = \int_0^1 \text{MTE}(u, X) \cdot (1\{p(X, \alpha_1(Z)) \geq u\} - 1\{p(X, \alpha_0(Z)) \geq u\})\, du$ for brevity.

Finally, we demonstrate that our binary formulation of encouragement rules nests treatment assignment rules that directly assign individuals to a certain treatment status as a special case. Suppose that α_0 and α_1 satisfy

$$p(X, \alpha_1(Z)) = 1 \text{ and } p(X, \alpha_0(Z)) = 0 \text{ almost surely}$$

so that $\alpha_1(Z)$ and $\alpha_0(Z)$ create perfectly strong incentives and disincentives to be in the treated state ($D = 1$), respectively, across heterogeneous covariate values. In this case, encouragement rules are effectively treatment assignment rules: $D(\boldsymbol{\alpha}^{\pi}(X, Z)) = \pi(X, Z)$.[4] Therefore, Corollary 2.2 provides a representation of the social welfare criterion for treatment assignment rules via the MTE function:

$$W(\boldsymbol{\alpha}^{\pi}) = E[Y(0)] + E\Big[\pi(X, Z) \int_0^1 \text{MTE}(u, X)\, du\Big].$$

[4] In this case, Z is redundant as a targeting variable.

This representation coincides with Theorem 1 of Sasaki and Ura (2024). However, such powerful manipulations are hard to justify in practice. For example, consider a selection of the form $D = 1\{Z \geq \tilde{U}\}$ with $\tilde{U}$ having full support on $\mathbb{R}$ and a manipulation of the form $\alpha_d(z) = z + a_d$ for $d = 0, 1$. Then, we need to set $a_1 = \infty$ and $a_0 = -\infty$ to induce full compliance.

2.3 Applications to EWM and Regret Properties

In this section, we restrict our attention to binary encouragement rules described in Section 2.2.3. We apply the representation of the social welfare criterion in Corollary 2.2 to the EWM framework and investigate the theoretical properties of the resulting statistical decision rules.

Some extra notations are needed to facilitate the discussion. Suppose the policymaker observes a random sample $A_i = (Y_i, D_i, X_i, Z_i)$ of size n. Let E_n denote the sample average operator, i.e., $E_n f = \frac{1}{n}\sum_{i=1}^n f(A_i)$ for any measurable function f. Let $a \vee b = \max\{a, b\}$.

For notational simplicity, we denote an encouragement rule and its social welfare by π and $W(\pi)$ in place of $\boldsymbol{\alpha}^\pi$ and $W(\boldsymbol{\alpha}^\pi)$, respectively. Let Π denote the class of encouragement rules the policymaker can choose from. Define the welfare contrast relative to the baseline policy that allocates everyone to α_0 as

$$\bar{W}(\pi) = E\Big[\pi(X, Z) \int_{p(X,\alpha_0(Z))}^{p(X,\alpha_1(Z))} \mathrm{MTE}(u, X)\, du\Big] = W(\pi) - E[Y(D(\alpha_0(Z)))].$$

By Corollary 2.2, the optimal encouragement rule is $\pi^* \in \arg\max_{\pi\in\Pi} \bar{W}(\pi)$ if the distribution of (X, Z) and the mappings $(u, x) \mapsto \mathrm{MTE}(u, x)$ and $(x, z) \mapsto p(x, z)$ are known. However, these quantities are unknown in practice. Corollary 2.2 implies that $\bar{W}(\pi)$ is point-identified under Assumption 2.2. Given an estimator $\hat{p}(x, z)$ for

$p(x,z)$ and an estimator $\widehat{\text{MTE}}(u,x)$ for $\text{MTE}(u,x)$, we construct the empirical welfare criterion $\hat{W}_n(\pi)$ by plugging in these estimators:

$$\hat{W}_n(\pi) = E_n\Big[\pi(X,Z)\int_{\hat{p}(X,\alpha_0(Z))}^{\hat{p}(X,\alpha_1(Z))} \widehat{\text{MTE}}(u,X)\,du\Big].$$

Then, we define the *feasible EWM encouragement rule* as

$$\hat{\pi}_{\text{FEWM}} \in \arg\max_{\pi\in\Pi} \hat{W}_n(\pi). \tag{2.3}$$

In line with the literature on statistical treatment rules (Manski, 2004; Kitagawa and Tetenov, 2018b), we evaluate the performance of an encouragement rule π by its *regret* defined as the welfare loss relative to the highest attainable welfare within class Π:

$$R(\pi) = \max_{\pi'\in\Pi} W(\pi') - W(\pi).$$

To analyze the regret of $\hat{\pi}_{\text{FEWM}}$, we impose the following assumptions.

Assumption 2.3 (Boundedness and VC-Class)**.** *(i) There exists $\bar{M} < \infty$ such that $\sup_{(u,x)\in[0,1]\times\mathcal{X}} |\,\text{MTE}(u,x)| \leq \bar{M}$. (ii) Π has a finite VC-dimension.*

Assumption 2.3(i) requires the MTE to be uniformly bounded in u and x. Assumption 2.3(ii) controls the complexity of the class Π of candidate encouragement rules in terms of VC-dimension. Interested readers can refer to van der Vaart and Wellner (1996) for the definition and textbook treatment of VC-dimension. We now give two examples of Π that satisfy Assumption 2.3(ii).

Example 2.4 (Linear Eligibility Score)**.** *Let $v \in \mathbb{R}^{d_v}$ be a subvector of (x,z). Consider the class of binary decision rules based on linear eligibility scores:*

$$\Pi_{\text{LES}} = \{\pi : \pi(x,z) = 1\{\lambda_0 + \lambda^\top v \geq 0\}, (\lambda_0,\lambda^\top) \in \mathbb{R}^{d_v+1}\}.$$

For example, individuals are assigned scholarships if a linear function of their tuition fee and distance to school exceeds some threshold. The EWM method searches over

all possible linear coefficients. The VC-dimension of Π_{LES} *is* $d_v + 1$.

Example 2.5 (Threshold Allocations)**.** *Consider the class of binary decision rules based on threshold allocations:*

$$\Pi_{\text{TA}} = \{\pi : \pi(x,z) = 1\{\sigma_k v_k \leq \bar{v}_k \text{ for } k \in \{1,\ldots,d_v\}\}, \bar{v} \in \mathbb{R}^{d_v}, \sigma \in \{-1,1\}^{d_v}\}.$$

For example, individuals are assigned scholarships if their tuition fee and distance to school are above or below some thresholds. The EWM method searches over all possible thresholds and directions. The VC-dimension of Π_{TA} *is* d_v.

We also need the following assumption about the estimators $\hat{p}(x,z)$ and $\widehat{\text{MTE}}(u,x)$. For brevity, let $\bar{\mathcal{U}}(x) = \cup_{d\in\{0,1\}}[\min \mathcal{P}_{\alpha_d}(x), \max \mathcal{P}_{\alpha_d}(x)]$, where $\mathcal{P}_{\alpha_d}(x)$ denotes the support of $p(X, \alpha_d(Z))$ conditional on $X = x$. As evident from Assumption 2.2(ii), point-identification of $\text{MTE}(\cdot, x)$ over $\bar{\mathcal{U}}(x)$ for every $x \in \mathcal{X}$ is one of the two key conditions that give rise to point-identification of $W(\pi)$.

Assumption 2.4 (Estimation of the Propensity Score and the MTE)**.**

(i) $E[E_n[\sup_{u\in\bar{\mathcal{U}}(X)} |\widehat{\text{MTE}}(u,X)|^2]] = O(1)$.

(ii) There exists a sequence $\psi_n \to \infty$ *such that for each* $d \in \{0,1\}$,

$$E[E_n[|\hat{p}(X,\alpha_d(Z)) - p(X,\alpha_d(Z))|^2]]^{1/2} = O(\psi_n^{-1}).$$

(iii) There exists a sequence $\phi_n \to \infty$ *such that*

$$E\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)} |\widehat{\text{MTE}}(u,X) - \text{MTE}(u,X)|\Big]\Big] = O(\phi_n^{-1}).$$

Assumption 2.4(i) concerns the second moment of $\widehat{\text{MTE}}(u,x)$. Assumptions 2.4(ii) and (iii) concern the convergence rate in expectation of the average estimation error for $\hat{p}(x,z)$ and $\widehat{\text{MTE}}(u,x)$. Since U is not observed, we take the supremum over the region where the MTE is point-identified. When the propensity score is estimated nonparametrically as in Example 2.1, a sufficient condition for Assumption 2.4(ii) is

$E[\sup_{(x,z)\in \text{Supp}(X,Z)} |\hat{p}(x,z) - p(x,z)|^2]^{1/2} = O(\psi_n^{-1})$. In Appendix B.4, we derive the sup-norm convergence rate in expectation for local polynomial estimators and series estimators built on exponential tail bounds. The rate can be faster if a semiparametric or parametric estimator is used under additional assumptions as in Example 2.2 or 2.3. The MTE is usually estimated using the estimated propensity score as a generated regressor. When the regression model is parametric, we provide sufficient conditions for Assumption 2.4(iii) to hold with $\phi_n = \psi_n$ in Appendix B.5. When the regression model is nonparametric, the sup-norm convergence rate in probability is established in Mammen et al. (2012, Corollary 1). However, the sup-norm convergence rate in expectation remains unknown. We leave it for future work.

Theorem 2.2 derives a convergence rate upper bound for the average regret of $\hat{\pi}_{\text{FEWM}}$. A proof is provided in Appendix B.1. If some nonparametric or semiparametric estimator is used for $\hat{p}(x,z)$ and $\widehat{\text{MTE}}(u,x)$, ψ_n^{-1} and ϕ_n^{-1} are slower than $n^{-1/2}$ in general. Hence, the rate upper bound is determined by $\psi_n^{-1} \vee \phi_n^{-1}$. In Appendix B.6, we pursue a doubly robust approach in the spirit of Athey and Wager (2021) to achieve $1/\sqrt{n}$-bounds for regret.

Theorem 2.2. *Under Assumptions 2.1-2.4, we have*

$$E[R(\hat{\pi}_{\text{FEWM}})] = O(\psi_n^{-1} \vee \phi_n^{-1} \vee n^{-1/2}).$$

2.4 Extensions

We consider two empirically relevant extensions to the baseline setup in Section 2.2.1. In Section 2.4.1, we allow for the presence of other instruments in addition to the one that can be manipulated. In Section 2.4.2, we incorporate budget constraints. As a further extension, we consider encouragement rules with a binary instrument in Appendix B.3.

2.4.1 Multiple Instruments

In practice, the policymaker can observe multiple instruments, but only one of them can be used as the tool for policy intervention. For example, tuition subsidies and proximity to upper secondary schools are two instruments for enrollment in upper secondary school, but only the former can serve as an encouragement. More generally, we allow Z to be L-dimensional with support $\mathcal{Z} \subset \mathbb{R}^L$. For each $z \in \mathcal{Z}$ and $\ell = 1, \ldots, L$, let z_ℓ denote the ℓth component of z, and let $z_{-\ell}$ collect all other $(L-1)$ components. Let Z_1 be the instrument that can be intervened upon. An encouragement rule is a mapping $\boldsymbol{\alpha} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ that determines the manipulated level of Z_1 while leaving Z_{-1} unchanged.

For each $\ell = 1, \ldots, L$ and $z_\ell \in \mathcal{Z}_\ell$, define the *marginal potential treatment* as the potential treatment status if Z_ℓ were set to z_ℓ while $Z_{-\ell}$ remained at its observed realization: $D_\ell(z_\ell) \equiv D(z_\ell, Z_{-\ell})$. We construct a selection equation for each instrument as

$$D_\ell(z_\ell) = 1\{p(X, z_\ell, Z_{-\ell}) \geq U_\ell\} \quad \text{with} \quad U_\ell | X, Z \sim \text{Unif}[0,1], \tag{2.4}$$

where U_ℓ can be interpreted as a latent proneness to take the treatment, which is measured against the incentive (or disincentive) created by the ℓth instrument. By (2.4), we only impose restrictions along each margin of selection and thus are agnostic about unobserved heterogeneity in the marginal rate of substitution across instruments.[5] Chen and Xie (2022) adhere to (2.2) when they deal with multiple instruments, thereby presenting the same social welfare representation as in the single-instrument case (i.e., Theorem 2.1).

[5] Mogstad et al. (2021) use a random utility model to demonstrate that in the presence of multiple instruments, (2.2) implies homogeneity in the marginal rate of substitution. In contrast, (2.4) does not impose such implicit homogeneity.

The outcome that would be observed under encouragement rule $\boldsymbol{\alpha}$ is

$$Y(D_1(\boldsymbol{\alpha}(X,Z))) = Y(1)\cdot D_1(\boldsymbol{\alpha}(X,Z)) + Y(0)\cdot(1 - D_1(\boldsymbol{\alpha}(X,Z))).$$

Define the social welfare criterion as $W(\boldsymbol{\alpha}) = E[Y(D_1(\boldsymbol{\alpha}(X,Z)))]$. Since (2.4) implies exogeneity for Z_ℓ in the sense of $\{D_\ell(z_\ell)\}_{z_\ell\in\mathcal{Z}_\ell} \perp Z_\ell | X, Z_{-\ell}$, where $\mathcal{Z}_\ell$ denotes the support of Z_ℓ, we accordingly replace Assumption 2.1(ii) with an exclusion restriction for Z_ℓ as follows:

Assumption 2.5 (Instrument-Specific Exclusion Restriction)**.** *For each $\ell = 1,\dots,L$, $E[Y(d)|X,Z,U_\ell] = E[Y(d)|X,Z_{-\ell},U_\ell]$ and $E[|Y(d)|] < \infty$ for $d \in \{0,1\}$.*

It turns out that $W(\boldsymbol{\alpha})$ can be expressed as a function of the instrument-specific MTE defined as

$$\text{MTE}_\ell(u_\ell, x, z_{-\ell}) \equiv E[Y(1) - Y(0)|U_\ell = u_\ell, X = x, Z_{-\ell} = z_{-\ell}].$$

The expression is given in Corollary 2.3. The analysis in Section 2.3 then applies. Heuristically, MTE_ℓ is equivalent to the MTE function using the ℓth instrument and conditioning on the other instruments as covariates.

Corollary 2.3. *Under (2.4) and Assumption 2.5, the social welfare criterion for a given encouragement rule $\boldsymbol{\alpha}$ is given by*

$$W(\boldsymbol{\alpha}) = E[Y] + E\Big[\int_0^1 \text{MTE}_1(u_1, X, Z_{-1}) \cdot (1\{p(X,\boldsymbol{\alpha}(X,Z),Z_{-1}) \geq u_1\} - 1\{p(X,Z) \geq u_1\})\, du_1\Big].$$

2.4.2 Budget Constraints

Manipulating the instrument can be costly, especially when the instrument is a monetary variable such as price. In practice, the policymaker often faces budget constraints and wants to prioritize encouragement for the individuals who will benefit the most.

Incorporating budget constraints is of particular interest when the treatment effect is intrinsically positive. For example, Dupas (2014) documents an experiment in Kenya that randomly assigned subsidized prices for a new health product. The treatment and outcome were indicators for the product's purchase and usage, respectively. The product was not available outside the experiment, so the potential outcome if not treated is identically equal to zero. Hence, the first-best decision rule was to assign the treatment, or an encouragement that induced one-way flows into treatment, to everyone. However, to preserve financial resources, in this scenario, the policymaker may wish to exclude individuals who are not likely to increase product usage, for example, because of low disease risks in their neighborhood.

Let $C : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_+$ be a user-chosen cost function that potentially depends on $\boldsymbol{\alpha}$. For example, $C(X, Z) = |\boldsymbol{\alpha}(X, Z) - Z|$ is a direct measure of manipulation costs.[6] For encouragement rule $\boldsymbol{\alpha}$, we define its budget by aggregating the costs for individuals who actually take up the treatment: $B(\boldsymbol{\alpha}) = E[C(X, Z) \cdot D(\boldsymbol{\alpha}(X, Z))]$. We consider settings in which the policymaker faces a harsh budget constraint such that the cost of implementing any encouragement rule cannot exceed κ.

Remark 2.1. *The policymaker may only want to account for cost without imposing a fixed budget, which is the thought experiment considered in Kitagawa and Tetenov (2018b) and Chen and Xie (2022). In this case, one can redefine the social welfare criterion as $W(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha})$ to apply the analysis in Section 2.3.*

As in Section 2.3, we specialize to binary encouragement rules when discussing the performance of statistical decision rules and denote the cost function by $B(\pi)$ in place of $B(\boldsymbol{\alpha}^\pi)$. Given a class Π of feasible encouragement rules,[7] the policymaker

[6] Depending on the context, additional costs can be embedded in the experimental design. For example, in the experiment documented in Thornton (2008), besides the monetary incentives for learning HIV results (the instrument), there were considerably high costs for testing, counseling/giving results, and selling condoms (see Table 12).

[7] We implicitly assume that there exists $\pi \in \Pi$ such that $B(\pi) \leq \kappa$.

now solves a constrained optimization problem:

$$\max_{\pi \in \Pi} W(\pi) \text{ s.t. } B(\pi) \leq \kappa.$$

Let π_{B}^* denote the oracle solution. It is helpful to introduce two desirable properties for statistical decision rules in the current setting: *asymptotic optimality* and *asymptotic feasibility*. Intuitively, with a large enough sample size, asymptotic optimality imposes that a statistical decision rule $\hat{\pi}$ is unlikely to achieve strictly lower welfare than π_{B}^*, and asymptotic feasibility imposes that $\hat{\pi}$ is unlikely to strictly violate the budget constraint.

Definition 2.1. *A statistical decision rule $\hat{\pi}$ is* asymptotically optimal *if, for any $\epsilon > 0$,*

$$\limsup_{n \to \infty} \Pr(W(\hat{\pi}) - W(\pi_{\mathrm{B}}^*) < -\epsilon) = 0.$$

A statistical decision rule $\hat{\pi}$ is asymptotically feasible *if, for any $\epsilon > 0$,*

$$\limsup_{n \to \infty} \Pr(B(\hat{\pi}) - \kappa > \epsilon) = 0.$$

Remark 2.2. *Asymptotic optimality and asymptotic feasibility are also considered in Sun (2024) but are defined asymmetrically. On one hand, asymptotic optimality only requires the population welfare of a statistical decision rule to concentrate around the optimal value from below. On the other hand, asymptotic feasibility requires the statistical decision rule to satisfy the population budget constraint without any slackness and thus is extremely sensitive to sampling uncertainty. In consequence, Sun (2024) proves the negative result that no statistical decision rule can uniformly satisfy both properties over a sufficiently rich class of data generating processes. In contrast, we show that it is possible to construct a statistical decision rule that simultaneously achieves both properties.*

Note that by (2.2),

$$B(\pi) = E[C(X, Z) \cdot \{\pi(X, Z) \cdot p(X, \alpha_1(Z)) + (1 - \pi(X, Z)) \cdot p(X, \alpha_0(Z))\}].$$

Define the *budget-constrained EWM encouragement rule* defined as a solution to the

sample version of the population constrained optimization problem:

$$\hat{\pi}_{\text{BEWM}} \in \arg\max_{\pi \in \Pi} \hat{W}_n(\pi) \text{ s.t. } \hat{B}_n(\pi) \leq \kappa, \tag{2.5}$$

where

$$\hat{B}_n(\pi) = E_n[C(X,Z) \cdot \{\pi(X,Z) \cdot \hat{p}(X, \alpha_1(Z)) + (1 - \pi(X,Z)) \cdot \hat{p}(X, \alpha_0(Z))\}].$$

We set $\hat{\pi}_{\text{BEWM}} = \emptyset$ if no $\pi \in \Pi$ satisfies $\hat{B}_n(\pi) \leq \kappa$. Theorem 2.3 asserts that $\hat{\pi}_{\text{BEWM}}$ satisfies both properties in Definition 2.1. A proof is provided in Appendix B.1.

Theorem 2.3. *Suppose that Assumptions 2.1–2.4 hold, and that the cost function $C(x,z)$ is uniformly bounded in x and z. Then, $\hat{\pi}_{\text{BEWM}}$ is asymptotically optimal and asymptotically feasible.*

2.5 Empirical Application

In this section, we apply the feasible EWM encouragement rule and the budget-constrained EWM encouragement rule to provide guidance on how to encourage upper secondary schooling, using data from the third wave of the Indonesian Family Life Survey (IFLS) fielded from June through November 2000. Carneiro et al. (2017) used this dataset to study the returns to upper secondary schooling in Indonesia. We follow Carneiro et al. (2017) in restricting our sample to males aged 25–60 who are employed and who have non-missing reported wage and schooling information. This subsample consists of 2,104 individuals.[8]

We specify the relevant variables in our framework as follows. The outcome Y is the log of hourly wages constructed from self-reported monthly wages and hours worked per week. The treatment D is an indicator of attendance of upper secondary

[8]The subsample used in Carneiro et al. (2017) does not contain the tuition fee variable, which plays a central role in our framework as the manipulatable instrument. Hence, we followed their descriptions to construct our subsample from raw data downloaded from the RAND Corporation website.

school or higher, corresponding to 10 or more years of completed education. The first instrument Z_1 is the lowest fee per continuing student, in thousands of rupiah, among secondary schools in the community of current residence.[9] The second instrument Z_2 is the distance, in kilometers, from the office of the community head of current residence to the nearest secondary school, which we define as the secondary school closest to the office of the community head.[10] We treat Z_1 as manipulatable and Z_2 as not manipulatable. Collect $Z = (Z_1, Z_2)^\top$. The covariates X include age, age squared, an indicator of rural residence, distance from the office of the community head of residence to the nearest health post, and indicators for religion, parental education, and the province of residence. Table B.2 in Appendix B.8 presents sample averages of these variables.

We focus on binary encouragement rules that manipulate Z_1 according to $\boldsymbol{\alpha}^\pi(X, Z) = \pi(X, Z) \cdot \alpha_1(Z_1) + (1 - \pi(X, Z)) \cdot \alpha_0(Z_1)$. For the binary decision π, we consider the class of linear rules based on (Z_1, Z_2):

$$\Pi_{\text{LES}} = \big\{\pi : \pi(x, z) = 1\{\lambda_0 + \lambda_1 \cdot z_1 + \lambda_2 \cdot z_2 > 0\}, \lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}\big\}.$$

We specify the manipulation function as $\alpha_1(Z_1) = (Z_1 - a) \cdot 1\{Z_1 \geq a\}$ and $\alpha_0(Z_1) = Z_1$. Here, $\alpha_1(Z_1)$ describes a tuition subsidy of up to a and $\alpha_0(Z_1)$ describes the status quo. The policymaker *a priori* chooses from $a \in \{2.5, 22.25\}$, which correspond to the sample median and maximum of Z_1, respectively. We specify the cost function

[9] The term "community" refers to the lowest-level administrative division in Indonesia. A community can either be a *desa* (village) or a *kelurahan* (urban community).

[10] The validity of an instrument constructed in this way can be controversial. Each individual's tuition and proximity are based on their current residence rather than their residence at the time of the secondary schooling decision. Educated individuals may move to more urban areas with more schools and higher tuition fees. Nonetheless, we note that the instrumental variable independence assumption for unrestricted instruments has testable implications, which are the generalized instrumental inequalities proposed by Kédagni and Mourifié (2020). Their tests suggest that the independence assumption between individuals' potential earnings and college tuition at age 17 is not rejected by the data from the 1979 National Longitudinal Survey of Youth used in Heckman et al. (2001).

as $C(X,Z) = |\boldsymbol{\alpha}^{\pi}(X,Z) - Z_1|$ and the budget constraint as $\kappa = 0.28$, which is about one-tenth of the average hourly wage.

The fact that Z_1 has discrete support (with 56 distinct values) violates the support condition for nonparametric or semiparametric identification of the propensity score.[11] Therefore, we estimate the propensity score from a logit regression of D on X, Z, $Z_1 \cdot Z_2$, and interactions between Z and X.[12] Although all elements of X are discrete, they together provide sufficient variation in the propensity score for the semiparametric estimation of the MTE.[13] We specify the conditional mean of Y given X, Z_2, and $p(X,Z)$ as

$$E[Y|X = x, Z_2 = z_2, p(X,Z) = u] = u(x^{\top}, z_2)\beta_1 + (1-u)(x^{\top}, z_2)\beta_0 + G(u),$$

where $G(\cdot)$ is an unknown function. By Corollary 2.3, the social welfare criterion of encouragement rule π is identified as $W(\pi) = E[Y] + E[\pi(X,Z) \cdot ((p(X, \alpha_1(Z_1), Z_2) - p(X,Z))(X^{\top}, Z_2)(\beta_1 - \beta_0) + G(p(X, \alpha_1(Z_1), Z_2)) - G(p(X,Z)))]$. We use the double residual regression procedure of Robinson (1988) to estimate (β_1, β_0). Given the estimators $\hat{p}(x,z)$ and $(\hat{\beta}_1, \hat{\beta}_0)$, we estimate $G(\cdot)$ using a nonparametric regression of the residual $Y - \hat{p}(X,Z)(X^{\top}, Z_2)\hat{\beta}_1 - (1 - \hat{p}(X,Z))(X^{\top}, Z_2)\hat{\beta}_0$ on $\hat{p}(X,Z)$. We use the locally linear regression throughout with a Gaussian kernel and a bandwidth of 0.06, which is determined by leave-one-out cross-validation.

We compute the feasible EWM encouragement rule $\hat{\pi}_{\text{FEWM}}$ in (2.3) and the budget-constrained EWM encouragement rule $\hat{\pi}_{\text{BEWM}}$ in (2.5) using the CPLEX mixed integer optimizer. Based on the decomposition result in Corollary 2.1, we report in Table 2.1 the estimated welfare gains, the average change in treatment take-up, and

[11] When $a = 2.5$, only 20 out of the 35 support points of $\alpha_1(Z_1)$ lie in the support of Z_1. When $a = 22.25$, $\alpha_1(Z_1)$ is identically equal to 0, which falls outside the support of Z_1.

[12] This specification of propensity score is an adaption of that considered by Carneiro et al. (2017) and Sasaki and Ura (2021), who use a single instrument Z_2.

[13] The estimated propensity score takes 1,782 distinct values that almost cover the full unit interval.

the PRTE of alternative encouragement rules, as well as the estimated proportion of eligible individuals. The PRTE measures the average change in the log of hourly wages among those induced to enroll in or drop out of upper secondary school. The seemingly favorable tuition subsidy has little effect on overall upper secondary school attendance when applied to everyone, resulting in a welfare gain of only a small magnitude. In contrast, the feasible EWM encouragement rule and the budget-constrained EWM encouragement rule achieve higher welfare gains by targeting a subpopulation with both a greater increase in treatment take-up and higher PRTE.

Table 2.1: Estimated Welfare Gain of Alternative Encouragement Rules

Policy	Share of Eligible Population	Est. Welfare Gain	Avg. Change in Treatment Take-up	PRTE
Panel A: $a = 2.5$ *(tuition subsidy up to the median tuition fee)*				
$\hat{\pi}_{\text{FEWM}}$	0.388	0.0146	0.0173	0.843
$\hat{\pi}_{\text{BEWM}}$ $(\kappa = 0.28)$	0.280	0.0102	0.0137	0.743
$\pi(x, z) = 1\ \forall x, z$	1	0.0005	0.0022	0.230
Panel B: $a = 22.25$ *(full tuition waiver)*				
$\hat{\pi}_{\text{FEWM}}$	0.386	0.0218	0.0317	0.688
$\hat{\pi}_{\text{BEWM}}$ $(\kappa = 0.28)$	0.286	0.0090	0.0141	0.639
$\pi(x, z) = 1\ \forall x, z$	1	0.0044	0.0140	0.315

We plot the feasible EWM encouragement rule and the budget-constrained EWM encouragement rule in Panels A and B of Figure 2·1, respectively. The shaded areas indicate the subpopulations to whom the tuition subsidy should be assigned. For both subsidy levels, the feasible EWM encouragement rule gives eligibility to individuals facing relatively high tuition fees and living relatively close to the nearest secondary school. The subpopulations targeted by the budget-constrained EWM encouragement rule shrink to the left. When the subsidy level a is increased from 2.5 to 22.25, the budget-constrained EWM encouragement rule tends to prioritize individuals facing

relatively low tuition fees.

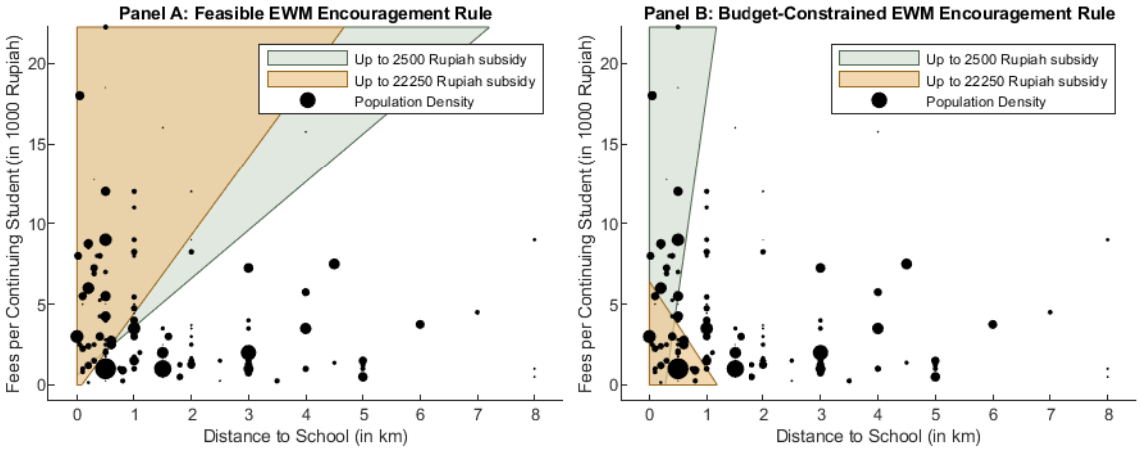


Figure 2·1: Targeted Subpopulation under Alternative Encouragement Rules

Utilizing a decomposition of the social welfare criterion analogous to Corollary 2.1, we offer a partial explanation of why the subpopulations indicated by the shaded areas in Figure 2·1 are targeted. We can write

$$W(\pi) = E[Y] + E[\pi(X,Z) \cdot \text{PRTE}(X,Z) \cdot (p(X,\alpha_1(Z_1),Z_2) - p(X,Z))],$$

where

$$\text{PRTE}(x,z) = \frac{\int_{p(x,z)}^{p(x,\alpha_1(z_1),z_2)} \text{MTE}_1(u_1,x,z_2)\,du_1}{p(x,\alpha_1(z_1),z_2) - p(x,z)}$$

measures the average treatment effect among individuals with $(X,Z) = (x,z)$ who are induced to switch treatment status when going from the status quo to a tuition subsidy. Let $\text{Med}(X)$ denote the sample median of X. We focus on the case of $a = 22.25$, i.e., a full tuition waiver. Panel A of Figure 2·2 displays the level sets of $(z_1,z_2) \mapsto \hat{p}(\text{Med}(X),\alpha_1(z_1),z_2) - \hat{p}(\text{Med}(X),z)$, namely the changes in treatment take-up for individuals with different values of (Z_1,Z_2) and the median value of X. We observe that only individuals with low Z_2 are induced into treatment. Panel B of

Figure 2·2 displays the level sets of

$$(z_1, z_2) \mapsto \widehat{\text{PRTE}}(\text{Med}(X), z) = \frac{\int_{\hat{p}(\text{Med}(X),z)}^{\hat{p}(\text{Med}(X),\alpha_1(z_1),z_2)} \widehat{\text{MTE}}_1(u_1, \text{Med}(X), z_2)\, du_1}{\hat{p}(\text{Med}(X), \alpha_1(z_1), z_2) - \hat{p}(\text{Med}(X), z)}.$$

We note that individuals induced into treatment have positive PRTE except for those with low values of (Z_1, Z_2). Put together, absent budget constraints, individuals in the upper-left corner are prioritized for the full tuition waiver. Meanwhile, contour lines where Z_1 is high are steep in both panels, implying that individuals with higher Z_1 incur greater manipulation costs for the same amount of welfare gains. Consequently, the optimal policy under budget constraints, which trades off welfare gains against costs, gives up individuals with high Z_1.

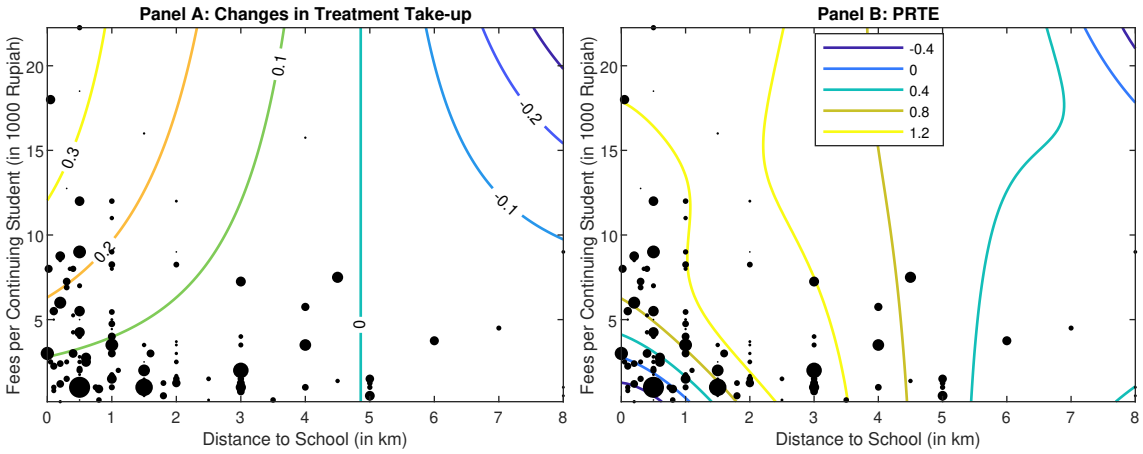


Figure 2·2: Impact of Going from the Status Quo to a Full Tuition Waiver

Notes: Panel A displays the level sets of $(z_1, z_2) \mapsto \hat{p}(\text{Med}(X), \alpha_1(z_1), z_2) - \hat{p}(\text{Med}(X), z)$. Panel B displays the level sets of $(z_1, z_2) \mapsto \widehat{\text{PRTE}}(\text{Med}(X), z)$. The size of the black dots indicates the number of individuals with different values of (Z_1, Z_2).

2.6 Conclusion

In this chapter, we propose a framework for policy learning that allows for endogenous treatment selection by leveraging an instrumental variable. To deal with imperfect compliance when designing policies, we consider encouragement rules in addition to

treatment assignment rules. To deal with failure of unconfoundedness when identifying the social welfare criterion, we incorporate the MTE function. We apply the representation of the social welfare criterion of binary encouragement rules via the MTE to the EWM method and derive convergence rates of regret. We also consider extensions allowing for multiple instruments and budget constraints. We illustrate the EWM encouragement rule using data from the Indonesian Family Life Survey. To be clear, the analysis in this chapter critically relies on the point-identification of the social welfare criterion. The necessary support condition or parametric assumptions could be restrictive. An interesting avenue for future research is to incorporate approaches to policy learning under partial identification of policy parameters (e.g., Russell (2020), D'Adamo (2022), Christensen et al. (2023), Yata (2025)).

Chapter 3

Model Selection Tests for Incomplete Models

3.1 Introduction

Selecting a suitable model is a crucial step in many empirical studies. Since the work of Akaike (1973), the Kullback-Leibler information criterion (KLIC) has been a central tool to measure a model's fit to the unknown data-generating process (DGP). When two competing models are considered, a natural way to select one of them is to compare their fit by the difference between their KLIC. Since the seminal work of Vuong (1989), model selection tests based on the sample counterpart of this quantity, the log-likelihood ratio (LR), have been applied widely.[1]

The goal of this chapter is to expand the scope of the likelihood-based model selection tests to a wide range of discrete choice models. Discrete choice models describe how an outcome Y is generated from economic primitives and observable covariates X. These models are commonly used to analyze economic decisions, such as household consumption, labor supply, firm entry, and government regulatory choices. In recent applications, models with set-valued (or incomplete) predictions have become more common because of their flexibility to accommodate strategic interaction, dynamic behavior, and rich unobserved heterogeneity. Examples include discrete

[1] See, Fafchamps (1993), Palfrey and Prisbrey (1997), Cameron and Heckman (1998), Caballero and Engel (1999), Nyarko and Schotter (2002), Coate and Conlin (2004), Paulson et al. (2006), Barseghyan et al. (2013), Francois et al. (2015), Kendall et al. (2015), to name a few.

games (Ciliberto and Tamer, 2009), dynamic discrete choice models (Honoré and Tamer, 2006; Berry and Compiani, 2022; Chesher et al., 2024), discrete choice models with heterogeneous choice sets (Barseghyan et al., 2021) or endogeneity (Chesher and Rosen, 2017), auctions under general bidding behavior (Haile and Tamer, 2003), network formation (Miyauchi, 2016; Sheng, 2020), product offerings (Eizenberg, 2014), exporter's decisions (Dickstein and Morales, 2018) and school choices (Fack et al., 2019). Incomplete models allow the model to predict multiple outcome values. The development of the applications above reflects the researchers' willingness to remain robust to certain aspects of their models that are not fully understood.

Just like conventional complete models, two different economic structures with incompleteness can lead to distinct predictions, resulting in different explanatory power regarding the observed data. In this context, practitioners face the choice of a model specification. For instance, when analyzing discrete games, one might want to compare a game of strategic substitution with a game of strategic complementarity. It is also common to compare a complete baseline model with a more general incomplete model that includes the former as a special case. For example, in the context of export decisions, Dickstein and Morales (2018) compared parameter estimates from a complete perfect foresight model with those from a more general model that relaxed the assumptions about the firms' information set.

While model selection is crucial in these situations, applying Vuong's (1989) original likelihood ratio test becomes challenging when at least one of the models is incomplete. This difficulty arises because an incomplete model can have multiple likelihoods for each parameter value. Additionally, the parameters in such models are often only partially identified. Therefore, any model selection test must address these non-standard features. These complexities may explain why a direct analogy to Vuong's test has not yet been developed.

We address these challenges by constructing a likelihood as follows. For each model and parameter θ, we consider a population problem of selecting the density $q_{\theta,y|x}$ closest in the KLIC to the DGP density $p_{0,y|x}$ among the ones consistent with θ. As shown in Kaido and Molinari (2024), finding such a density can be formulated as a convex program. Upon solving the problem, we impose the model's *sharp identifying restrictions* as constraints. This ensures that the likelihood uses all information in each model. We then form a likelihood ratio using the KLIC projection $q_{\theta,y|x}$ while replacing the unknown DGP $p_{0,y|x}$ with a nonparametric estimator $\hat{p}_{n,y|x}$. This construction generalizes the standard likelihood framework to incomplete models. Using the KLIC to construct a model density is in the spirit of Vuong (1989). We note that the models under consideration may be misspecified (White, 1994). Hence, the goal here is to select a model that is closer to the DGP in terms of the chosen information criterion.

We further study the asymptotic properties of the proposed LR statistic. Vuong (1989) demonstrated that the limiting distribution of the standard LR statistic changes, depending on whether the two models overlap or not. This feature also applies to our statistic, posing a challenge in ensuring the uniform validity of inference across different DGPs. To address this challenge, we incorporate regularization into the statistic based on the work of Shi (2015b) and Schennach and Wilhelm (2017). The regularization ensures that our proposed test statistic is asymptotically non-degenerate and follows a standard normal distribution, regardless of the underlying DGP, thus making inference more tractable. This tractability comes at the cost of choosing a regularization parameter. We examine how to choose its value through simulations.

This chapter contributes to the literature of model selection in parametric models that followed Vuong (1989). Rivers and Vuong (2002) consider model selection criteria other than the likelihood function to allow for a broad class of estimation methods

and dynamic models, with a focus on mean squared errors of prediction. Li (2009) employs simulated mean squared errors of prediction to deal with complex structural models. Chen et al. (2007) compare a parametric model with a moment equality model. Shi (2015b) and Schennach and Wilhelm (2017) modify the classical Vuong test to achieve uniform size control for overlapping and nonoverlapping models. Shi (2015b) uses the local asymptotic theory to design a higher-order bias correction and a variance adjustment to the test statistic. Liao and Shi (2020) extend this idea to semi/non-parametric models. Schennach and Wilhelm (2017) add noise to the test statistic by sample-splitting. However, none of the aforementioned tests can accommodate incomplete models.

To our knowledge, Shi (2015a) and Hsu and Shi (2017) are the only model selection tests that can accommodate an incomplete model. Their tests are based on the generalized empirical likelihood (GEL) statistic for models characterized by moment restrictions. Shi (2015a) only considers a finite number of unconditional moment restrictions. Hsu and Shi (2017) propose the average generalized empirical likelihood (AGEL) to handle conditional moment restrictions. Their tests are applicable if both competing models are characterized by moment equalities or inequalities. If any of the models are complete, one needs to derive equivalent conditional moment restrictions instead of directly using the likelihood. Our approach complements theirs by providing an alternative statistic that compares the likelihoods of the two models, which does not require the researcher to derive conditional moment restrictions. This approach is computationally tractable when likelihoods are available in closed form in many cases. The recent work of Chen and Kaido (2023) develops a score test for testing the null hypothesis of model completeness against an incomplete alternative. Our tests differ from theirs in that (i) the competing models can be incomplete in our setting, whereas one of the models must be complete and nested by the other model

in their setting; and (ii) the competing models can be misspecified in our framework.

Throughout, we use upper case letters (e.g., W) to represent a random element and lower case letters (e.g., w) to denote the specific values the random element can take. For any random elements A and B, $A \sim B$ means equality in distribution, and $A \perp B$ means statistical independence. We use $A|B$ to represent the conditional distribution of A given B. We write the support of A and the conditional support of A given B as $\text{supp}(A)$ and $\text{supp}(A|B)$, respectively.

3.2 Set-up and Notation

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ and $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ denote, respectively, observable endogenous and exogenous variables, and $U \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$ denote latent variables. We assume $\mathcal{Y}$ is a finite set. Let $P_0 \in \Delta(\mathcal{Y} \times \mathcal{X})$ denote the distribution of (Y, X), where for a space $\mathcal{S}$, $\Delta(\mathcal{S})$ denotes the set of all Borel probability measures on $(\mathcal{S}, \Sigma_{\mathcal{S}})$, and $\Sigma_{\mathcal{S}}$ is the Borel σ-algebra on $\mathcal{S}$. For $\mathcal{S} = \mathcal{Y} \times \mathcal{X}$ we let $\Sigma_{\mathcal{S}}$ equal the product σ-algebra $\Sigma_Y \times \Sigma_X$.

Suppose a model imposes restrictions on the joint behavior of (Y, X, U), and that these restrictions are expressed through a measurable correspondence known up to a finite-dimensional parameter vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$:

$$Y \in G(U|X;\theta), \ a.s. \tag{3.1}$$

If G is singleton-valued $a.s.$, there exists a function $g : \mathcal{U} \times \mathcal{X} \times \Theta \to \mathcal{Y}$ such that

$$Y = g(U|X;\theta). \tag{3.2}$$

The structure in (3.1) therefore nests standard complete discrete choice models.

Let $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ denote a family of distributions for the latent variables U, known up to a finite-dimensional parameter vector that is part of θ. An economic

structure is then summarized by the tuple $(G, \Theta, \mathcal{F})$. We illustrate the objects above with known examples. The first example is a discrete game of complete information (Bresnahan and Reiss, 1990; Tamer, 2003).

Example 3.1 (Discrete Games)**.** Consider a binary-response game with two players (e.g., firms). Each player may either choose $y^{(j)} = 0$ or $y^{(j)} = 1$. Player j's payoff is

$$\pi^{(j)} = y^{(j)}\big(x^{(j)\prime}\delta^{(j)} + \beta^{(j)}y^{(-j)} + u^{(j)}\big), \tag{3.3}$$

where $y^{(-j)} \in \{0, 1\}$ denotes the other player's action, $x^{(j)}$ is player j's observable characteristics, and $u^{(j)}$ is a payoff shifter that is unobservable to the econometrician. The payoff is assumed to belong to the players' common knowledge. A policy-relevant parameter is the *strategic interaction effect* $\beta^{(j)}$ which captures the impact of the opponent's taking $y^{(-j)} = 1$ on player j's payoff. The sign of this parameter determines the nature of the game.

Suppose the players play a pure strategy Nash equilibrium (PSNE), but the researcher does not know the equilibrium selection. Let $y = (y^{(1)}, y^{(2)})$. In the presence of negative externalities, i.e., $\beta^{(j)} < 0$ for $j = 1, 2$, one can summarize the set of PSNEs by the following correspondence (Beresteanu et al., 2011, Proposition 3.1), where $\theta = (\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)})$:

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in S_{\{(0,0)\}|x;\theta} \equiv \{u : u^{(j)} < -x^{(j)\prime}\delta^{(j)}, j = 1, 2\}, \\ \{(0,1)\} & u \in S_{\{(0,1)\}|x;\theta} \equiv \{u : u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}\}, \\ & \quad \cup\{u : -x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} > x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\} \\ \{(1,0)\} & u \in S_{\{(1,0)\}|x;\theta} \equiv \{u : u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ & \quad \cup\{u : -x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \\ \{(1,1)\} & u \in S_{\{(1,1)\}|x;\theta} \equiv \{u : u^{(j)} > -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, j = 1, 2\}, \\ \{(1,0),(0,1)\} & u \in S_{\{(0,1),(1,0)\}|x;\theta} \equiv \{u : -x^{(j)\prime}\delta^{(j)} < u^{(j)} < -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, \\ & \qquad j = 1, 2\}. \end{cases} \tag{3.4}$$

Note that the model predicts multiple equilibria when $u \in S_{\{(0,1),(1,0)\}|x;\theta}$ (see Figure 3·1). A special case of this model is Berry's (1992) specification that assumes $\delta^{(j)} = \delta$ and $\beta^{(j)} = \beta$ for all j. Under this symmetry assumption, the equilibrium number of entrants is uniquely determined. Let $N = y^{(1)} + y^{(2)}$ and let $\theta = (\beta, \delta)$. Then, the complete prediction of the model is

$$N = g(u|x;\theta) = \begin{cases} 0 & u \in S_{\{(0,0)\}|x;\theta} \\ 1 & u \in S_{\{(0,1)\}|x;\theta} \cup S_{\{(0,1),(1,0)\}|x;\theta} \cup S_{\{(1,1)\}|x;\theta} \\ 2 & u \in S_{\{(1,1)\}|x;\theta}. \end{cases}$$

Another possible structure to consider is a game of strategic complementarity, i.e., $\beta^{(j)} > 0, j = 1, 2$. This structure's predicted equilibria are

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in S_{\{(0,0)\}|x;\theta} \equiv \{u : u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \\ & \qquad \cup \{u : -x^{(1)\prime}\delta^{(1)} - \beta^{(1)} \leq u^{(1)} < -x^{(1)\prime}\delta^{(1)}, \\ & \qquad u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ \{(0,1)\} & u \in S_{\{(0,1)\}|x;\theta} \equiv \{u : u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} \geq -x^{(2)\prime}\delta^{(2)}\}, \\ \{(1,0)\} & u \in S_{\{(1,0)\}|x;\theta} \equiv \{u : u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, \\ & \qquad u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ \{(1,1)\} & u \in S_{\{(1,1)\}|x;\theta} \equiv \{u : u^{(1)} \geq -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, \\ & \qquad u^{(2)} \geq -x^{(2)\prime}\delta^{(2)}\} \\ & \qquad \cup \{u : u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, \\ & \qquad -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \leq u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}, \\ \{(0,0),(1,1)\} & u \in S_{\{(0,0),(1,1)\}|x;\theta} \equiv \{u : -x^{(j)\prime}\delta^{(j)} - \beta^{(j)} \leq u^{(j)} < -x^{(j)\prime}\delta^{(j)}, \\ & \qquad j = 1, 2\}. \end{cases} \tag{3.5}$$

In this case, the model predicts $(0, 0)$ and $(1, 1)$ as multiple equilibria for some value of u.

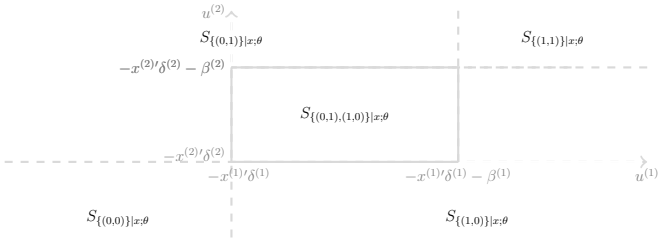


Figure 3·1: Level Sets of $G(\cdot|x;\theta)$ with $\beta^{(j)} < 0, j = 1, 2$.

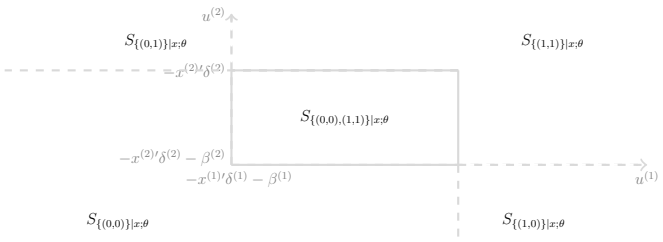


Figure 3·2: Level Sets of $G(\cdot|x;\theta)$ with $\beta^{(j)} > 0, j = 1, 2$.

The second example is a multinomial choice model with heterogeneous choice sets.

Example 3.2 (Heterogeneous Choice Sets)**.** Consider a discrete choice model, with a finite universe of alternatives $\mathcal{J} = \{1, \ldots, J\}$. Each alternative is characterized by a vector of covariates X_j, which might vary across decision makers, and let $X = [X_j, j \in \mathcal{J}]$. Let U denote a vector representing the individual's unobserved taste.

The decision maker faces a *choice set* $C \subseteq \mathcal{J}$ and chooses the alternative $Y \in C$ that maximizes their utility:

$$Y \in \arg\max_{j \in C} W(j, X, U; \theta). \tag{3.6}$$

The researcher observes (Y, X), but not C, and wishes to learn features of θ.

One may take different strategies to treat the choice set formation process. One possibility is to model the process explicitly. For example, Goeree (2008) specifies a parametric model of random choice sets. This approach determines the conditional distribution of $(C, U)|X$. The complete model in (3.6) then induces a unique likelihood function.[2] Alternatively, Barseghyan et al. (2021) only assume each decision

[2]This is provided that a tie occurs with probability 0 or the researcher specifies a tie-breaking

maker draws a set of cardinality at least κ. For given $\theta \in \Theta$ and $x \in \mathcal{X}$, Barseghyan et al. (2021, Lemma A.1) show that the set of optimal choices is a measurable correspondence:

$$G(U|x;\theta) = \cup_{K \subseteq \mathcal{J}:|K| \geq \kappa} \{\arg\max_{j \in K} W(j,x,U;\theta)\}. \tag{3.7}$$

3.2.1 Comparing Models

Let $\mathcal{C}$ be the collection of all subsets of $\mathcal{Y}$. Define the *containment functional* $\nu_\theta : \mathcal{C} \times \mathcal{X} \to [0,1]$ associated with G by

$$\nu_\theta(A|x) = \int_{\mathcal{U}} 1\{G(u|x;\theta) \subseteq A\} dF_\theta(u), \quad \forall A \in \mathcal{C}. \tag{3.8}$$

This functional uniquely determines the distribution of the random set $G|X$ (Molchanov, 2005). The model prediction (3.1), however, does not uniquely determine the conditional distribution of the outcome Y. Artstein's theorem (see e.g. Molinari, 2020) ensures the set of all model predicted conditional distributions of Y satisfying (3.1) is the *core* of ν_θ:

$$\text{core}(\nu_\theta(\cdot|x)) \equiv \{Q \in \mathcal{M}(\Sigma_Y, \mathcal{X}) : Q(A|x) \geq \nu_\theta(A|x), A \in \mathcal{C}\}. \tag{3.9}$$

A system of inequalities $Q(\cdot|x) \geq \nu_\theta(\cdot|x)$, called the *sharp identifying restrictions*, characterizes the core. They are known to contain all information in the underlying model (Galichon and Henry, 2011).

Assume that there are σ-finite measures μ and ν on $(\mathcal{Y}, \Sigma_Y)$ and $(\mathcal{X}, \Sigma_X)$, respectively, a product measure $\zeta \equiv \mu \times \nu$ on $(\mathcal{Y} \times \mathcal{X}, \Sigma_{YX})$, and that for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $Q \in \text{core}(\nu_\theta(\cdot|x))$, $Q \ll \mu$. Then, given (3.9), one can define the set of conditional

rule.

densities associated with core$(\nu_\theta(\cdot|x))$:

$$\mathfrak{q}_\theta \equiv \left\{ q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu,\ Q \in \text{core}(\nu_\theta(\cdot|x)),\ x \in \mathcal{X} \right\}. \tag{3.10}$$

Following Kaido and Molinari (2024), we define a *model* as the collection of sets $\mathfrak{q}_\theta$ across $\theta \in \Theta$:

$$\mathfrak{Q} \equiv \{\mathfrak{q}_\theta : \theta \in \Theta\}.$$

Consider competing structures $(G_s, \Theta_s, \mathcal{F}_s), s = 1, 2$. For each s, let $\mathfrak{Q}_s$ be the model induced by structure $(G_s, \Theta_s, \mathcal{F}_s)$. We compare the models in terms of their closeness to the true density $p_0 = dP_0/d\mu$. For a measure space $(\Omega, \mathfrak{F}, \zeta)$, let $f : \Omega \mapsto \mathbb{R}_+$ be a measurable function satisfying $\int f d\zeta < \infty$ and $\int_S f \ln f d\zeta < \infty$ where $S = \{\omega \in \Omega : f(\omega) > 0\}$. The *Kullback-Leibler Information Criterion* (KLIC) between f and another density f' is

$$I(f||f') \equiv \int_S f \ln \frac{f}{f'} d\zeta. \tag{3.11}$$

Let $\mathfrak{f}$ denote a *set* of measurable functions $f' : \Omega \mapsto \mathbb{R}_+$ satisfying $\int_S f \ln f' d\zeta < \infty$. The KLIC between f and $\mathfrak{f}$ is

$$I(f||\mathfrak{f}) \equiv \inf_{f' \in \mathfrak{f}} I(f||f') \tag{3.12}$$

Given a joint density function $f(y, x)$, its associated conditional density function $f(y|x)$, and another conditional density function $f'(y|x)$, we denote their conditional KLIC by

$$I(f||f') \equiv \int_{\mathcal{Y}\times\mathcal{X}} f(y, x) \ln \frac{f(y|x)}{f'(y|x)} d\zeta(y, x) \tag{3.13}$$

and use Eq. (3.13) in the KL divergence measure in Eq. (3.12).

We aim to test the following null hypothesis:

$$H_0 : I(p_0||\mathfrak{Q}_1) = I(p_0||\mathfrak{Q}_2). \tag{3.14}$$

It states that the two structures induce models that attain the same value of KLIC to p_0. A one-sided alternative hypothesis is

$$H_1 : I(p_0||\mathfrak{Q}_1) < I(p_0||\mathfrak{Q}_2). \tag{3.15}$$

One can select Model 1 over Model 2 if the test suggests strong evidence against H_0 in favor of H_1.

We recast the comparison of KLIC into a comparison of expected log-likelihood functions, building on the insights of Akaike (1973) and White (1994). For each model, $I(p_0||\mathfrak{Q}) = \inf_{\theta\in\Theta} I(p_0||\mathfrak{q}_\theta)$. For each θ, the following equalities hold:

$$\begin{aligned} I(p_0||\mathfrak{q}_\theta) &= \inf_{q\in\mathfrak{q}_\theta} \int_{\mathcal{Y}\times\mathcal{X}} p_0(y,x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\zeta(y,x) \\ &= \int_{\mathcal{X}} p_{0,x}(x) \inf_{q_{y|x}\in\mathfrak{q}_{\theta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y) d\nu(x), \end{aligned} \tag{3.16}$$

where

$$\mathfrak{q}_{\theta,x} = \{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu,\ Q \in \text{core}(\nu_\theta(\cdot|x))\}. \tag{3.17}$$

The inner optimization problem in (3.16) is a convex program with a strictly convex objective function. Hence, a unique solution exists. Let

$$q^*_{\vartheta,y|x}(\cdot|x; p_{0,y|x}) = \arg\min_{q_{y|x}\in\mathfrak{q}_{\vartheta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y). \tag{3.18}$$

One can view $q^*_{\vartheta,y|x}$ as the projection of p_0 on $\mathfrak{q}_\vartheta$ via KLIC. Since Y is discrete, it is straightforward to compute this solution by solving the convex program. We illustrate

how to derive this object in Section 3.4. Below, we call the map $\vartheta \to q^*_{\vartheta,y|x}$ *profiled-likelihood function* because it is a function obtained by profiling out the selection mechanism.[3] Below, for each s, let $q^*_{\theta_s,y|x}$ denote the profiled-likelihood in model s.

Recall that the KL divergence is $I(p_0||\mathfrak{q}_\theta) = E_{P_0}[\ln p_0(Y|X)] - E_{P_0}[\ln q^*_{\theta,y|x}(Y|X)]$ whose first term does not depend on the models. For each $s \in \{1,2\}$, denote the value function of the convex program associated with the KL divergence between $\mathfrak{q}_{\theta_s,x}$ and $p_{y|x}$ by

$$\begin{aligned} L(x,\theta_s,p_{y|x}) &\equiv \sup_{q_{y|x}\in\mathfrak{q}_{\theta_s,x}} \int_{\mathcal{Y}} p_{y|x}(y|x)\ln q_{y|x}(y|x;p_{y|x})d\mu(y) \\ &= E_P[\ln q^*_{\theta_s,y|x}(Y|X;p_{y|x})|X=x]. \end{aligned}$$

Then, we can reformulate the null hypothesis as

$$H_0 : E_{P_0}[L(X,\theta^*_1,p_{0,y|x})] = E_{P_0}[L(X,\theta^*_2,p_{0,y|x})],$$

where θ^*_s is a maximizer of $\theta_s \mapsto E_{P_0}[L(X,\theta_s,p_{0,y|x})], s = 1,2$.[4] This reformulation shows the model comparison boils down to comparing the maximized expected likelihoods, where each likelihood function is $q^*_{\vartheta,y|x}$. When the underlying structure is complete, $q^*_{\vartheta,y|x}$ coincides with the standard likelihood function because each $\mathfrak{q}_\theta$ is a singleton set. Hence, this construction nests Vuong's (1989) original framework as a special case.

We adopt the following definition of correct specification from Kaido and Molinari (2024).

Definition 3.1 (Correctly Specified Model & Misspecified Model)**.** *A model is correctly specified if $p_0 \in \mathfrak{q}_\theta$ for some $\mathfrak{q}_\theta \in \mathfrak{Q} \equiv \{\mathfrak{q}_\vartheta : \vartheta \in \Theta\}$, and misspecified otherwise.*

[3] This is because each element of $\text{core}(\nu_\theta(\cdot|x))$ can also be written as the set of probability measures such that $Q(\cdot|x) = \int \eta(\cdot|x,u)dF_\theta(u|x)$, where η is a conditional distribution (selection mechanism) supported on $G(u|x;\theta)$.

[4] The maximizer is not necessarily unique. This does not cause a problem because the results below does not require a unique maximizer of the objective function.

Hence, a model is misspecified when one cannot recover the DGP p_0 even if one augments the model by a selection mechanism. For conceptual purposes, it is useful to define how the two models relate to each other using the profiled likelihood. We follow Vuong (1989) and Liao and Shi (2020) to introduce the following terms.

Definition 3.2 (Strictly Non-nested Models)**.** *Models* $\mathfrak{Q}_1$ *and* $\mathfrak{Q}_2$ *are* strictly nonnested *if there does not exist* $(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ *such that* $q^*_{\theta_1,y|x}(y|x; p_{0,y|x}) = q^*_{\theta_2,y|x}(y|x; p_{0,y|x})$ *for all* $(y, x) \in \mathcal{Y} \times \mathcal{X}$*.*

Definition 3.3 (Overlapping Models)**.** *Models* $\mathfrak{Q}_1$ *and* $\mathfrak{Q}_2$ *are* overlapping *if they are not strictly nonnested.*

Definition 3.4 (Nested Models)**.** *Model* $\mathfrak{Q}_1$ nests $\mathfrak{Q}_2$ *if, for each* $\theta_2 \in \Theta_2$*, there exists* $\theta_1 \in \Theta_1$ *such that* $q^*_{\theta_1,y|x}(y|x; p_{0,y|x}) = q^*_{\theta_2,y|x}(y|x; p_{0,y|x})$ *for all* $(y, x) \in \mathcal{Y} \times \mathcal{X}$*.*

3.2.2 Test Statistic and Implementation

Our test uses a sample analog of the following quasi log-likelihood ratio (QLR)

$$\text{QLR}_{P_0} = \max_{\theta_1 \in \Theta_1} E_{P_0}[L(X, \theta_1, p_{0,y|x})] - \max_{\theta_2 \in \Theta_2} E_{P_0}[L(X, \theta_2, p_{0,y|x})],$$

and examines if it is far enough from 0.

Suppose a sample $\{(Y_i, X_i), i = 1, \ldots, n\}$ of size n is available. Our test adds suitable regularization to a QLR statistic so that it admits an asymptotically normal approximation over a wide class of DGPs. We will explain how cross-fitting and regularization ensure the uniform validity of inference in Section 3.3. Below, we first describe the algorithm to compute the test statistic.

Algorithm 1: (Cross-fit QLR-test):

Step 0: Split the entire sample (indexed by $i \in \{1, \ldots, n\}$) into two equal halves denoted by I_1 and I_2. For each $\ell \in \{1, 2\}$, let $I_{-\ell} = \{1, \ldots, n\} \setminus I_\ell$.

Step 1: For each $\ell \in \{1, 2\}$, estimate $p_{0,y|x}$ nonparametrically using the observations in I_ℓ; denote the resulting estimator by $\hat{p}_{I_\ell,y|x}$.

Step 2: For each $s, \ell \in \{1, 2\}$, compute the weighted quasi maximum-likelihood estimator (QMLE) of θ_s by plugging in $\hat{p}_{I_\ell, y|x}$ for p_0 and using the observations in I_ℓ:

$$\hat{\theta}_{s,I_\ell} \in \arg\max_{\theta_s \in \Theta_s} \frac{2}{n} \sum_{i \in I_\ell} L(X_i, \theta_s, \hat{p}_{I_\ell, y|x}).$$

Step 3: For each $\ell \in \{1, 2\}$, calculate the QLR statistic by plugging in $\hat{p}_{I_\ell, y|x}$ for p_0, $\hat{\theta}_{1,I_{-\ell}}$ for θ_1, and $\hat{\theta}_{2,I_{-\ell}}$ for θ_2, and using the observations in I_ℓ:

$$\widehat{\mathrm{QLR}}_{I_\ell} = \frac{2}{n} \sum_{i \in I_\ell} L(X_i, \hat{\theta}_{1,I_{-\ell}}, \hat{p}_{I_\ell, y|x}) - \frac{2}{n} \sum_{i \in I_\ell} L(X_i, \hat{\theta}_{2,I_{-\ell}}, \hat{p}_{I_\ell, y|x}).$$

Step 4: For each $\ell \in \{1, 2\}$, calculate the variance statistic by plugging in $\hat{p}_{I_\ell, y|x}$ for p_0, $\hat{\theta}_{1,I_{-\ell}}$ for θ_1, and $\hat{\theta}_{2,I_{-\ell}}$ for θ_2, and using the observations in I_ℓ:

$$\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = \hat{\sigma}^2_{1,I_\ell}(\hat{\theta}_{1,I_{-\ell}}) - 2\hat{\sigma}_{12,I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\sigma}^2_{2,I_\ell}(\hat{\theta}_{2,I_{-\ell}}),$$

where $\hat{\theta}_{I_\ell} = (\hat{\theta}'_{1,I_\ell}, \hat{\theta}'_{2,I_\ell})'$,

$$\begin{aligned}
\hat{\sigma}^2_{s,I_\ell}(\theta_s) &= \frac{2}{n} \sum_{i \in I_\ell} \Big(\ln q^*_{\theta_s, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) - \frac{2}{n} \sum_{i \in I_\ell} \ln q^*_{\theta_s, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) \Big)^2, \quad s = 1, 2 \\
\hat{\sigma}_{12,I_\ell}(\theta) &= \frac{2}{n} \sum_{i \in I_\ell} \Big(\ln q^*_{\theta_1, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) - \frac{2}{n} \sum_{i \in I_\ell} \ln q^*_{\theta_1, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) \Big) \\
&\quad \times \Big(\ln q^*_{\theta_2, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) - \frac{2}{n} \sum_{i \in I_\ell} \ln q^*_{\theta_2, y|x}(Y_i|X_i; \hat{p}_{I_\ell, y|x}) \Big).
\end{aligned}$$

Step 5: For each $\ell \in \{1, 2\}$, with an auxiliary random variable $U_\ell \sim N(0, 1)$ (independent of the sample), construct the subsample test statistic as

$$\hat{T}_{I_\ell} = \frac{\sqrt{n/2}\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}}, \tag{3.19}$$

where $\hat{\omega}_{I_\ell}$ is a data-dependent regularization parameter. We recommend

$$\hat{\omega}_{I_\ell} = (1 + C \cdot \ln n \cdot \hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}))^{-1} \tag{3.20}$$

for a user-chosen constant $C > 0$.

Step 6: Average across ℓ to obtain the final *cross-fit test statistic*:

$$\hat{T}_n = \frac{\hat{T}_{I_1} + \hat{T}_{I_2}}{\sqrt{2}}.$$

Let z_α denote the α quantile of $N(0,1)$. For the two-sided test, one can perform the following substeps. Namely, reject H_0 and pick model 1 if $\hat{T}_n > z_{1-\alpha/2}$, reject H_0 and pick model 2 if $\hat{T}_n < -z_{1-\alpha/2}$, and do not reject H_0 otherwise.

If one knows *a priori* $E_{P_0}[L(X, \theta^*_1, p_{0,y|x})] \geq E_{P_0}[L(X, \theta^*_2, p_{0,y|x})]$ one can conduct a one-sided test. For example, this approach can be used if the researcher knows Model 1 nests Model 2. Suppose the alternative hypothesis is

$$H_1 : E_{P_0}[L(X, \theta^*_1, p_{0,y|x})] > E_{P_0}[L(X, \theta^*_2, p_{0,y|x})]. \tag{3.21}$$

In this case, reject H_0 and pick model 1 if $\hat{T}_n > z_{1-\alpha}$, and do not reject H_0 otherwise.

The main component of the test statistic is the quasi-likelihood ratio $\widehat{\mathrm{QLR}}_{I_\ell}$, which compares the two models' fit to the data. A novel feature is that we use the profiled-likelihood $q^*_{\theta,y|x}$ to address the incompleteness of the model. The statistic has several additional features. First, it involves a regularization term $\hat{\omega}_{I_\ell} U_\ell$. This term keeps the statistic non-degenerate when $\widehat{\mathrm{QLR}}_{I_\ell}$'s variance is close to zero, a feature that is known to raise a challenge for uniformly valid inference when the two models overlap. Second, the denominator of $\hat{T}_{I_\ell}$ standardizes the statistic so that its asymptotic distribution is standard normal. Finally, we use the sample-splitting technique. We construct a parameter estimate $\hat{\theta}_{I_{-\ell}}$ from observations outside I_ℓ and

evaluate the likelihood on I_ℓ. This helps us ensure the validity of inference even when θ_s^* is only partially identified.

3.3 Asymptotic Properties of the QLR test

The QLR statistic outlined above is asymptotically normally distributed under H_0 and local alternatives. We provide high-level assumptions that ensure this result and use them to establish the asymptotic uniform validity of the test.

We first collect key objects for the theoretical analysis of the QLR statistic. For each s, define the *pseudo-true identified set* as the set of maximizers of the expected log-likelihood:

$$\Theta_s^*(p_0) \equiv \arg\max_{\theta_s \in \Theta_s} E_{P_0}[\ln q^*_{\theta_s,y|x}(Y|X; p_{0,y|x})].$$

This set collects the parameter values θ_s that minimize the KL divergence to the DGP p_0. It reduces to a singleton containing the *pseudo-true parameter value* θ_s^* as in White (1994) if the underlying structure is complete and θ_s^* is unique. If a model is correctly specified, θ_s^* coincides with the true value.

For each $s \in \{1, 2\}$, consider an arbitrary parameter value $\theta_s \in \Theta_s$. Define the projection of θ_s on $\Theta_s^*(P_0)$ by

$$\theta_s^*(\theta_s, P_0) = \arg\inf_{\theta_s^* \in \Theta_s^*(P_0)} \|\theta_s - \theta_s^*\|. \tag{3.22}$$

Let $\theta^*(\theta, P_0) = ((\theta_1^*(\theta_1, P_0))', (\theta_2^*(\theta_2, P_0))')'$.

Finally, for each function $f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, let $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n \big(f(Y_i, X_i) - E_{P_0}[f(Y_i, X_i)]\big)$. Let $\mathcal{H}$ be a parameter space to which $p_{0,y|x}$ belongs and let $\|p - p'\|_{\mathcal{H}}$ be a pseudo-metric on $\mathcal{H}$. For each $k = (k_1, k_2) \in \mathbb{N}^2$, let $\mathcal{F}^k = \{f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | f(y, x) = \prod_{s=1}^2 (\ln q^*_{\theta_s,y|x}(y|x; p_{y|x}))^{k_s},\ \theta \in \Theta,\ p_{y|x} \in \mathcal{H}\}$.

Throughout, we assume the availability of a random sample.

Assumption 3.1. $\{(Y_i, X_i)\}_{i=1}^n$ *are i.i.d. under* P_0.

We first analyze the asymptotic behavior of the QMLE $\hat{\theta}_{s,I_\ell}$ for $s, \ell \in \{1,2\}$. To characterize $\hat{\theta}_{s,I_\ell}$ and $\Theta_s^*(P_0)$ by first-order conditions, we make the following assumptions.

Assumption 3.2. *(a)* $\mathcal{Y}$ *is a finite set. For each* $s \in \{0,1\}$*, (b) there is a collection* $\mathcal{A}_G \subset 2^{\mathcal{Y}}$ *such that* $\mathcal{A}_G = \text{supp}(G(\cdot|x;\theta_s)) \equiv \{A \subseteq \mathcal{Y} : F_{\theta_s}(G(U|x;\theta_s) = A) > 0\}$ *for all* $\theta_s \in \Theta_s$ *and* $x \in \mathcal{X}$*. (c)* $\nu_{\theta_s}(A|x)$ *is continuously differentiable with respect to* θ_s *for all* $A \subset \mathcal{Y}$ *and* $x \in \mathcal{X}$*.*

Assumption 3.2(a) restricts attention to models with discrete outcomes. Assumption 3.2(b) requires the support of the correspondence $G(\cdot|x;\theta_s)$ not to vary with $\theta_s \in \Theta_s$. Assumption 3.2(c) is easily verified when F_{θ_s} is differentiable in θ_s. Under Assumption 3.2, we can establish the differentiability of $L(x,\theta_s,p_{y|x})$ with respect to θ_s (see Lemma C.1(i) in Appendix C). Then, $m(x,\theta_s,p_{y|x}) \equiv \frac{\partial}{\partial\theta_s}L(x,\theta_s,p_{y|x})$ is well-defined, and

$$\frac{2}{n}\sum_{i\in I_\ell} m(X_i,\hat{\theta}_{s,I_\ell},\hat{p}_{I_\ell,y|x}) = 0, \quad \ell = 1,2,$$

$$E_{P_0}[m(X,\theta_s^*,p_{0,y|x})] = 0 \quad \forall\theta_s^* \in \Theta_s^*(P_0).$$

We add the following regularity conditions to $m(x,\theta_s,p_{y|x})$. Hence, θ_s^* is characterized as a solution to the *score equation*, the system of equations defined by the expected value of $m(X,\cdot,p_{0,y|x})$. Similarly, $\hat{\theta}_{s,I_\ell}$ solves the sample analog of the score equation.

We add the following regularity conditions to $m(x,\theta_s,p_{y|x})$.

Assumption 3.3. *For each* $s \in \{0,1\}$*, (a) there exist positive constants* C *and* δ *such that*

$$\|E_{P_0}[m(X,\theta_s,p_{0,y|x})]\| \geq C \cdot (d(\theta_s,\Theta_s^*(P_0)) \wedge \delta),$$

where $d(a,B) \equiv \inf_{b\in B}\|a-b\|$*; (b)* $\sup_{\theta_s,p_{y|x}}\|\mathbb{G}_n(m(\cdot,\theta_s,p_{y|x}))\| = O_p(1)$*; (c) there*

exists a positive constant K_m *such that for any* $\theta_s \in \Theta_s$ *and* $p_{y|x}, \tilde{p}_{y|x} \in \mathcal{H}$,

$$\|E_{P_0}[m(X, \theta_s, p_{y|x})] - E_{P_0}[m(X, \theta_s, \tilde{p}_{y|x})]\| \leq K_m \|p_{y|x} - \tilde{p}_{y|x}\|_{\mathcal{H}}.$$

Assumption 3.4. *For each* $\ell \in \{0, 1\}$, $\|\hat{p}_{I_\ell, y|x} - p_{0,y|x}\|_{\mathcal{H}} = O_p(n^{-d_p})$ *for* $1/4 < d_p \leq 1/2$.

Assumption 3.3(a) ensures that $\theta_s \mapsto \|E_{P_0}[m(X, \theta_s, p_{0,y|x})]\|$ increases not too slowly as θ_s moves away from $\Theta_s^*(P_0)$. This is a high-level condition which needs to be checked in each example. Similar conditions are imposed in moment inequality models (Chernozhukov et al., 2007). Kaido et al. (2022) further discusses this type of condition and how to check them in specific examples. Assumption 3.3(b) requires the empirical process $\mathbb{G}_n(m(\cdot, \theta_s, p_{y|x}))$ to be stochastically bounded over $(\theta_s, p_{y|x})$. Assumption 3.3(c) imposes Lipschitz continuity on $E_{P_0}[m(X, \theta_s, p_{y|x})]$ with respect to $p_{y|x}$. Assumption 3.4 is a rate condition on $\hat{p}_{I_\ell, y|x}$, which can be satisfied by kernel and sieve estimators under suitable smoothness conditions on $p_{0,y|x}$. Under Assumptions 3.1–3.4, we may ensure that the QMLE $\hat{\theta}_{s,I_\ell}$ is in an n^{-d_p}-neighborhood of $\Theta_s^*(P_0)$ (see Lemma C.2 in Appendix C).

Next, we analyze the asymptotic behavior of the subsample QLR statistic $\widehat{\mathrm{QLR}}_{I_\ell}$ for $\ell \in \{1, 2\}$. Under Assumption 3.2, we can also establish the directional differentiability of $p_{y|x} \mapsto L(x, \theta_s, p_{y|x})$ with the directional derivative at $p_{y|x}$ in the direction $\tilde{p}_{y|x} - p_{y|x}$ denoted by $D(x, \theta_s, p_{y|x}, \tilde{p}_{y|x} - p_{y|x})$ (see Lemma C.1(ii) in Appendix C). We add the following regularity conditions.

Assumption 3.5. *(a) There exists a function* $B(\cdot) : \mathcal{X} \to \mathbb{R}$ *such that* $E[B^2(X)] < \infty$ *and for each* $s \in \{1, 2\}$,

$$\begin{aligned} \sup_{\theta_s \in \Theta_s, p_{y|x} \in \mathcal{H}} \|m(x, \theta_s, p_{y|x})\| &\leq B(x), \\ \sup_{\theta_s \in \Theta_s, p_{y|x}, \tilde{p}_{y|x} \in \mathcal{H}} |D(x, \theta_s, p_{y|x}, \tilde{p}_{y|x} - p_{y|x})| &\leq B(x), \end{aligned}$$

and for any $\theta_s^* \in \Theta_s^*(P_0)$ *and any* $\theta_s \in \Theta_s$ *and* $p_{y|x} \in \mathcal{H}$ *with* $\|\theta_s - \theta_s^*\|$ *and* $\|p_{y|x} -$

$p_{0,y|x}\|_{\mathcal{H}}$ *small enough,*

$$|L(x,\theta_s,p_{y|x}) - L(x,\theta_s^*,p_{0,y|x}) - (\theta_s-\theta_s^*)'m(x,\theta_s^*,p_{0,y|x}) - D(x,\theta_s^*,p_{0,y|x},p_{y|x}-p_{0,y|x})| \\ \leq B(x)(\|\theta_s-\theta_s^*\|^2 + \|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}^2).$$

(b) $\int_0^\infty \sqrt{\ln N(\varepsilon,\mathcal{H},\|\cdot\|_{\mathcal{H}})}d\varepsilon < \infty$*, where* $N(\varepsilon,\mathcal{H},\|\cdot\|_{\mathcal{H}})$ *denotes the covering number of size* ε *for* $\mathcal{H}$*, and for each* $s\in\{1,2\}$*,* Θ_s *is a compact subset of* $\mathbb{R}^{d_\theta}$*. (c) For each* $s\in\{1,2\}$*, for any* $\theta_s^*\in\Theta_s^*(P_0)$ *and* $p_{y|x}\in\mathcal{H}$ *with* $\|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}$ *small enough,*

$$\sqrt{n/2}E_{P_0}[D(X,\theta_s^*,p_{0,y|x},p_{y|x}-p_{0,y|x})] - \sqrt{2/n}\sum_{i\in I_\ell}\alpha(Y_i,X_i) = o_p(1),$$

where $\alpha(y,x)\equiv\sum_{\tilde{y}\in\mathcal{Y}}(1\{y=\tilde{y}\}-p_{0,y|x}(\tilde{y}|x))\ln q^*_{\theta_s^*,y|x}(\tilde{y}|x,p_{0,y|x})$.

Assumption 3.5(a) requires a dominance condition on $m(x,\theta_s,p_{y|x})$ and $D(x,\theta_s,p_{y|x},\tilde{p}_{y|x}-p_{y|x})$ and assumes that $L(X,\theta_s,p_{y|x})$ can be linearized in θ_s and $p_{y|x}$. Assumption 3.5(b) restricts the covering numbers for the parameter class $\{\theta_s\in\Theta_s, p_{y|x}\in\mathcal{H}: \|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}\leq\delta_n\}$. Assumption 3.5(c) is a simplified version of Assumption 5.3 of Newey (1994). It imposes the asymptotic equivalence between $E_{P_0}[D(X,\theta_s^*,p_{0,y|x},p_{y|x}-p_{0,y|x})]$ and the sample average of $\alpha(Y,X)$, where $\alpha(y,x)$ corresponds to the correction term for estimation of $p_{0,y|x}$ as characterized in Proposition 4 of Newey (1994).

Under Assumptions 3.1–3.5, we can obtain an asymptotically linear representation of each model's contribution to the subsample QLR statistic (see Lemma C.3 in Appendix C):

$$(n/2)^{-1/2}\sum_{i\in I_\ell}L(X_i,\hat{\theta}_{s,I_{-\ell}},\hat{p}_{I_\ell,y|x}) = (n/2)^{-1/2}\sum_{i\in I_\ell}\ln q^*_{\theta_s^*(\hat{\theta}_{s,I_{-\ell}},P_0),y|x}(Y_i|X_i;p_{0,y|x}) + o_p(1). \tag{3.23}$$

For brevity, let $\lambda_\theta(y|x;p_{y|x})$ denote the logarithm of the ratio of the two profiled-likelihood functions:

$$\lambda_\theta(y|x;p_{y|x}) = \ln q^*_{\theta_1,y|x}(y|x;p_{y|x}) - \ln q^*_{\theta_2,y|x}(y|x;p_{y|x}).$$

Applying the asymptotically linear representation in (3.23) to the two models, we may approximate the subsample QLR statistic as follows[5]

$$\begin{aligned}&\sqrt{n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_0}) \\ &= \sqrt{2/n}\sum_{i\in I_\ell}(\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}},P_0)}(Y_i|X_i;p_{0,y|x}) - E_{P_0}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}},P_0)}(Y|X;p_{0,y|x})]) + o_p(1). \quad (3.24)\end{aligned}$$

To ensure the asymptotic normality of the leading term in (3.24), we impose the following dominance condition on $\lambda_{\theta^*}(y|x;p_{0,y|x})$ for $\theta^* \in \Theta^*(P_0)$, which allows us to invoke Lyapounov's central limit theorem (see Lemma C.4 in Appendix C). Let $\sigma^2_{P_0}(\theta) = E_{P_0}[\lambda^2_\theta(Y|X;p_{0,y|x})] - E_{P_0}[\lambda_\theta(Y|X;p_{0,y|x})]^2$.

Assumption 3.6. *There exist positive constants M and ϵ and a function $D : \mathcal{Y}\times\mathcal{X} \to \mathbb{R}$ such that $E_{P_0}[|D(Y,X)|^{2+\epsilon}] \le M$ and for all $(y,x) \in \mathcal{Y}\times\mathcal{X}$ and $\theta^* \in \Theta^*(P_0)$, $|\lambda_{\theta^*}(y|x;p_{0,y|x}) - E_{P_0}[\lambda_{\theta^*}(Y|X;p_{0,y|x})]| \le D(y,x)\sigma_{P_0}(\theta^*)$.*

The asymptotic variance of the leading term in (3.24) is $\sigma^2_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}},P_0))$. We impose the following conditions for the estimation of $\sigma^2_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}},P_0))$.

Assumption 3.7. *(a) For each $s \in \{1,2\}$, $\sup_{\theta_s\in\Theta_s, p_{y|x}\in\mathcal{H}} E_{P_0}[|\ln q^*_{\theta_s,y|x}(Y|X;p_{y|x})|] = O_p(1)$. (b) For each $k=(k_1,k_2)\in\mathbb{N}^2$ such that $k_1+k_2\le 2$, $\sup_{f\in\mathcal{F}^k}|\mathbb{G}_n(f)| = O_p(1)$, where $\mathcal{F}^k \equiv \{f : \mathcal{Y}\times\mathcal{X}\to\mathbb{R} | f(y,x) = \prod_{s=1}^2(\ln q^*_{\theta_s,y|x}(y|x;p_{y|x}))^{k_s},\ \theta\in\Theta,\ p_{y|x}\in\mathcal{H}\}$. (b) There exists a positive constant K_λ such that for any $p_{y|x},\tilde{p}_{y|x}\in\mathcal{H}$ and $\theta,\tilde{\theta}\in\Theta$, $|E_{P_0}[\lambda^k_\theta(Y|X;p_{y|x})] - E_{P_0}[\lambda^k_{\tilde{\theta}}(Y|X;\tilde{p}_{y|x})]| \le K_\lambda(\|\theta-\tilde{\theta}\| + \|p_{y|x}-\tilde{p}_{y|x}\|_{\mathcal{H}})$, $k=1,2$.*

Assumption 3.7(a) bounds the first moment of the profiled log-likelihood. Assumption 3.7(b) assumes the maximum of an empirical process defined over $\mathcal{F}^k$ is stochastically bounded, which can be shown by applying a maximal inequality. Assumption 3.7(c) imposes Lipschitz continuity on $E_{P_0}[\lambda^k_\theta(Y|X;p_{y|x})]$ with respect to θ and $p_{y|x}$. Under Assumptions 3.1–3.7, we can show that $\sigma^2_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}},P_0))$ can be estimated at the same rate as $p_{0,y|x}$ using the subsample variance statistic $\hat{\sigma}_{I_\ell}(\hat{\theta}_{I_{-\ell}})$ (see Lemma C.5 in Appendix C).

[5]To be precise, we apply the result to a subsequence of DGPs along which the asymptotic size is attained.

A well-known issue with model selection tests is the possible degeneracy of the QLR statistic (Vuong, 1989; Shi, 2015b; Schennach and Wilhelm, 2017). In our context, $\sigma_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_0))$ can be arbitrarily close to 0. As a result, the leading term may not dominate the remainder term in (3.24), causing the asymptotic distribution of $\sqrt{n/2}(\widehat{\text{QLR}}_{I_\ell} - \text{QLR}_{P_0})$ to be non-normal. We restrict attention to DGPs under which $\sigma_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_0))$ only converges to zero at a polynomial rate.

Definition 3.5. *Let $\mathcal{P}$ be the set of DGPs such that Assumptions 3.1–3.7 hold and for each $\ell \in \{1,2\}$ and any sequence $\{P_n \in \mathcal{P}\}$, either $\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)) = O_p(n^{-d_\sigma})$ for some $d_\sigma > 0$ or $\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)) \xrightarrow{p} \sigma_\infty > 0$.*

The subsample test statistic in (3.19) adds a regularization term $\hat{\omega}_{I_\ell} U_\ell$, which keeps the statistic non-degenerate even if the subsample QLR statistic is degenerate. The recommended regularization parameter sequence in (3.20) has the following property (see Lemma C.6 in Appendix C):[6]

Condition 3.1. *For each $\ell \in \{1,2\}$ and any sequence $\{P_n \in \mathcal{P}\}$ such that $\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)) \xrightarrow{p} \sigma_\infty \in [0,\infty)$, (a) if $\sigma_\infty = 0$, we have $\hat{\omega}_{I_\ell} \xrightarrow{p} \omega_\infty > 0$; (b) if $\sigma_\infty > 0$, we have $\hat{\omega}_{I_\ell} \xrightarrow{p} 0$.*

Define $\mathcal{P}_0 \equiv \{P \in \mathcal{P} : E_P[L(X, \theta_1^*, p_{y|x})] = E_P[L(X, \theta_2^*, p_{y|x})]\}$. This is the subset of distributions in $\mathcal{P}$ that satisfy $H_0 : \text{QLR}_{P_0} = 0$. Define the two-sided model-selection test of level α as

$$\varphi_n^{\text{2-sided}}(\alpha) = 1\{|\hat{T}_n| > z_{1-\alpha/2}\},$$

and the one-sided model selection test of level α for H_0 against $H_1 : \text{QLR}_{P_0} > 0$ as

$$\varphi_n^{\text{1-sided}}(\alpha) = 1\{\hat{T}_n > z_{1-\alpha}\}.$$

The following theorem asserts that the proposed test achieves uniform asymptotic size control.

[6] Any other sequences that satisfy Condition 3.1 can be used.

Theorem 3.1. *Suppose Assumptions 3.1–3.7 hold. Then, for any sequence* $\{P_n \in \mathcal{P}_0\}$,

$$\lim_{n\to\infty} E_{P_n}[\varphi_n(\alpha)] = \alpha \ ,$$

for $\varphi_n = \varphi_n^{2\text{-sided}}$ *or* $\varphi_n = \varphi_n^{1\text{-sided}}$.

The following theorem characterizes the lower bound of the asymptotic power of the proposed test against local alternatives.

Theorem 3.2. *Suppose Assumptions 3.1–3.7 hold. Then, for any sequence* $\{P_n \in \mathcal{P} \setminus \mathcal{P}_0\}$ *such that for each* $\ell \in \{1,2\}$, $\sigma^2_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)) \xrightarrow{p} \sigma_\infty \in [0,\infty)$ *and* $\sqrt{n}\,\mathrm{QLR}_{P_n} \to h \in (0,\infty)$,

$$\liminf_{n\to\infty} E_{P_n}[\varphi_n^{2\text{-sided}}(\alpha)] \geq 1 - \Phi(z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)) + \Phi(-z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)),$$
$$\liminf_{n\to\infty} E_{P_n}[\varphi_n^{1\text{-sided}}(\alpha)] \geq 1 - \Phi(z_{1-\alpha} - h/(\omega_\infty \vee \sigma_\infty)).$$

3.4 Examples

We revisit the examples to illustrate the proposed test. For notational simplicity, we drop subscript s from the objects below.

3.4.1 Discrete games

For $s = 1$, the structure represents a game of strategic substitution. Let $\Theta = \{\theta = (\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}) : \beta^{(j)} \leq 0,\ \delta^{(j)} \in \Theta_\delta \subset \mathbb{R}^{d_\delta},\ j = 1, 2, \}$. The equilibrium correspondence G is as in (3.4). Suppose the distribution of $U = (U^{(1)}, U^{(2)})'$ belongs to a parametric family $\mathcal{F} = \{F_\theta(\cdot|x) : \theta \in \Theta\}$.

Define

$$\eta^1(\theta; x) \equiv F_\theta(S_{\{(1,0)\}|x;\theta}|x) + F_\theta(S_{\{(0,1),(1,0)\}|x;\theta}|x) + F_\theta(S_{\{(0,1)\}|x;\theta}|x)$$
$$\eta^2(\theta; x) \equiv F_\theta(S_{\{(1,0)\}|x;\theta}|x) + F_\theta(S_{\{(0,1),(1,0)\}|x;\theta}|x)$$
$$\eta^3(\theta; x) \equiv F_\theta(S_{\{(1,0)\}|x;\theta}|x).$$

Here, $\eta^1(\theta;x)$ is the predicted probability of either $Y=(1,0)$ or $(0,1)$. Similarly, $\eta^2(\theta;x)$ is the upper bound on the probability of $Y=(1,0)$, and $\eta^3(\theta;x)$ the lower bound on the probability of the same event (see Figure 3·1).

Let $p_{0,M}((1,0)|x) \equiv \frac{p_0((1,0)|x)}{p_0((1,0)|x)+p_0((0,1)|x)}$ be the relative frequency of outcome $(1,0)$ out of the "Monopoly" event $Y \in \{(1,0),(0,1)\}$. The profiled likelihood takes the following closed-form[7]:

$$q^*_{\theta,y|x}((0,0)|x;p_{0,y|x}) = F_\theta(S_{\{(0,0)\}|x;\theta}|x) \tag{3.25}$$

$$q^*_{\theta,y|x}((1,1)|x;p_{0,y|x}) = F_\theta(S_{\{(1,1)\}|x;\theta}|x) \tag{3.26}$$

$$q^*_{\theta,y|x}((1,0)|x;p_{0,y|x}) = p_{0,M}((1,0)|x)\eta^1(\theta;x)\mathbb{I}^1(x;\theta) + \eta^2(\theta;x)\mathbb{I}^2(x;\theta) + \eta^3(\theta;x)\mathbb{I}^3(x;\theta), \tag{3.27}$$

where

$$\mathbb{I}^1(x;\theta) \equiv 1\{\eta^3_1(\theta;x)/\eta^1(\theta;x) \le p_{0,M}((1,0)|x) \le \eta^2(\theta;x)/\eta^1(\theta;x)\}$$
$$\mathbb{I}^2(x;\theta) \equiv 1\{p_{0,M}((1,0)|x) > \eta^2(\theta;x)/\eta^1(\theta;x)\}$$
$$\mathbb{I}^3(x;\theta) \equiv 1\{p_{0,M}((1,0)|x) < \eta^3(\theta;x)/\eta^1(\theta;x)\}.$$

Let us explain the intuition behind (3.25)–(3.27). First, $q^*_{\theta,y|x}((0,0)|x;p_{0,y|x})$ is simply the probability allocated to $S_{\{(0,0)\}|x;\theta}$ because $Y=(0,0)$ is the unique equilibrium when $U \in S_{\{(0,0)\}|x;\theta}$. A similar argument applies to $q^*_{\theta,y|x}((1,1)|x;p_{0,y|x})$. Second, $q^*_{\theta,y|x}((1,0)|x;p_{0,y|x})$ depends on the relative frequency $p_{0,M}((1,0)|x)$. For each θ, the model predicts the relative frequency would lie in the interval $[\eta^3(\theta;x)/\eta^1(\theta;x), \eta^2(\theta;x)/\eta^1(\theta;x)]$. If $p_{0,M}((1,0)|x)$ is in the interval (case 1), the profiled likelihood is proportional to it. If $p_{0,M}((1,0)|x)$ is above $\eta^2(\theta;x)/\eta^1(\theta;x)$ (cases 2), the profiled likelihood in (3.27) is given by the upper bound $\eta^2(\theta;x)$ of the predicted probability

[7]See Kaido and Molinari (2024) for derivation. Since $\mathcal{Y}$ consists of four outcomes, we report the value of the profiled likelihood for three outcome values below.

of $Y = (1,0)$. Finally, if $p_{0,M}((1,0)|x)$ is below $\eta^3(\theta;x)/\eta^1(\theta;x)$ (case 3), the profiled likelihood in (3.27) is given by the lower bound $\eta^3(\theta;x)$ of the probability of $Y = (1,0)$.

For an alternative model ($s = 2$), consider one of Berry's (1992) specification which imposes the symmetry restriction $\beta^{(j)} = \beta$ and $\delta^{(j)} = \delta$ for $j = 1, 2$. This specification gives the following likelihood function for $N \in \{0,1,2\}$:

$$q^*_{\theta,y|x}(0|x;p_{0,y|x}) = F_\theta(S_{\{(0,0)\}|x;\theta}|x), \tag{3.28}$$

$$q^*_{\theta,y|x}(1|x;p_{0,y|x}) = F_\theta(S_{\{(0,1)\}|x;\theta}|x) + F_\theta(S_{\{(0,1),(1,0)\}|x;\theta}|x) + F_\theta(S_{\{(1,0)\}|x;\theta}|x), \tag{3.29}$$

$$q^*_{\theta,y|x}(2|x;p_{0,y|x}) = F_\theta(S_{\{(1,1)\}|x;\theta}|x), \tag{3.30}$$

where $\theta = (\beta, \delta)$.

As another competing model, consider a game of strategic complementarity. Let $\Theta_2 = \{\theta = (\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}) : \beta^{(j)} \geq 0,\ \delta^{(j)} \in \Theta_\delta \subset \mathbb{R}^{d_\delta},\ j = 1,2,\}$. The equilibrium correspondence for this case is as in (3.5). Let $p_{0,N}((1,1)|x) \equiv \frac{p_0((1,1)|x)}{p_0((0,0)|x)+p_0((1,1)|x)}$ be the relative frequencies of outcome $(1,1)$ out of the "Non-Monopoly" event $Y \in \{(0,0),(1,1)\}$. An argument similar to structure 1 shows the profiled-likelihood is

$$q^*_{\theta,y|x}((1,0)|x;p_{0,y|x}) = F_\theta(S_{\{(1,0)\}|x;\theta}|x) \tag{3.31}$$

$$q^*_{\theta,y|x}((0,1)|x;p_{0,y|x}) = F_\theta(S_{\{(0,1)\}|x;\theta}|x) \tag{3.32}$$

$$\begin{aligned} q^*_{\theta,y|x}((0,0)|x;p_{0,y|x}) = {} & p_{0,N}((0,0)|x)\eta^1(\theta_2;x)\mathbb{I}^1(x;\theta) + [\eta^1(\theta;x) - \eta^2(\theta;x)]\mathbb{I}^2(x;\theta) \\ & + [\eta^1(\theta;x) - \eta^3(\theta;x)]\mathbb{I}^3(x;\theta), \end{aligned} \tag{3.33}$$

where

$$\begin{aligned} \eta^1(\theta;x) &= F_\theta(S_{\{(1,1)\}|x;\theta}|x) + F_\theta(S_{\{(0,0),(1,1)\}|x;\theta}|x) + F_\theta(S_{\{(0,0)\}|x;\theta}|x) \\ \eta^2(\theta;x) &= F_\theta(S_{\{(1,1)\}|x;\theta}|x) + F_\theta(S_{\{(0,0),(1,1)\}|x;\theta}|x) \\ \eta^3(\theta;x) &= F_\theta(S_{\{(1,1)\}|x;\theta}|x), \end{aligned}$$

and

$$\mathbb{I}^1(x;\theta) \equiv 1\{\eta^3(\theta;x)/\eta^1(\theta;x) \leq p_{0,N}((1,1)|x) \leq \eta^2(\theta;x)/\eta^1(\theta;x)\}$$
$$\mathbb{I}^2(x;\theta) \equiv 1\{p_{0,N}((1,1)|x) > \eta^2(\theta;x)/\eta^1(\theta;x)\}$$
$$\mathbb{I}^3(x;\theta) \equiv 1\{p_{0,N}((1,1)|x) < \eta^3(\theta;x)/\eta^1(\theta;x)\}.$$

3.4.2 Heterogeneous Choice Sets

Consider a choice of insurance plans. An individual faces a risk of a loss that occurs with probability $\mu \in [0,1]$. Insurance plans $\{1,\ldots,J\}$ are available. Each plan is characterized by the deductible c_j and insurance premium π_j and defines a binary lottery $L_j = (-\pi_j, 1-\mu; -\pi_j - c_j, \mu)$. Let $v(\cdot;U)$ be the von-Neumann Morgenstern utility function with risk aversion coefficient U. The risk aversion coefficient is unknown to the econometrician and is assumed to follow a distribution $F_{U|X,\theta}$. For each j, define

$$W(L_j;U) = \mu v(-\pi_j - c_j;U) + (1-\mu)v(-\pi_j;U). \tag{3.34}$$

Each individual chooses a plan that maximizes the expected utility $W(\cdot;U)$ from a choice set $C \subset \{1,\ldots,J\}$. The observable outcome is the selected plan $Y \in \{1,\ldots,J\}$ and individual characteristics $X_j = (c_j, \pi_j, \mu)$. These variables can be used to make inference for the risk preference (Cohen and Einav, 2007; Barseghyan et al., 2011).

The first structure specifies the unobserved choice set's conditional distribution following Goeree (2008). Suppose C and U are independent conditional on X. For any $K \subset \{1,\ldots,J\}$, the conditional probability of $C = K$ is

$$F_{C|X,\theta}(K|x) \equiv P(C = K|X = x) = \prod_{l\in K} \phi_l(x) \prod_{k\notin K} (1-\phi_k(x)), \tag{3.35}$$

where $\phi_l(x) = \frac{\exp(x'\gamma_l)}{1+\exp(x'\gamma_l)}$ is the probability that the individual becomes aware of

alternative l (e.g., thorough advertisement).[8] Let

$$f_\theta(j|x,K) = \int 1\{W(L_j;u) > W(L_k;u), \forall k \in K, k \neq j\} dF_{U|X,\theta}(u) \tag{3.36}$$

It represents the conditional probability of the agent choosing plan $j \in K$ given $(X,C) = (x,K)$. Let $\mathcal{C}_j$ be the set of all choice sets containing product j. The structure above is complete and induces the following likelihood function:

$$q^*_{\theta_1,y|x}(j|x) = \sum_{K \in \mathcal{C}_j} \prod_{l \in K} \phi_l(x) \prod_{k \notin K} (1 - \phi_k(x)) f_\theta(j|x,K), \;\; j = 1, \ldots, J. \tag{3.37}$$

In a competing model, we allow C and U to be related arbitrarily. Following Barseghyan et al. (2021), we assume C contains at least κ elements, which induces (3.7). In what follows, we assume there are low, medium, and high deductible plans such that $c_1 < c_2 < c_3$. A low deductible means higher coverage since it ensures lower out-of-pocket payments when a loss occurs. Accordingly, the insurance premia are assumed to satisfy $\pi_j = b_j \pi$ with $b_1 > b_2 > b_3$, where π is an individual-specific base price. Suppose v belongs to the family of utility functions with negligible third derivative (NTD), i.e.,

$$\frac{v(w+\Delta)}{v'(w)} - \frac{v(w)}{v'(w)} = \Delta - \frac{U}{2}\Delta^2. \tag{3.38}$$

Then, there exists a threshold $\bar{\tau}(x)$ of U at which the individual is indifferent between plans 1 and 3 (Barseghyan et al., 2021). If (and only if) the agent's risk aversion is below the threshold, less coverage is always preferred to more coverage for all U, i.e., $3 \succsim 2 \succsim 1$. In contrast, if $U \geq \bar{\tau}(x)$, we have the opposite ordering.

Suppose $\kappa = 2$. Then, possible choice sets are $\{1,2\}, \{2,3\}, \{1,3\}$, and $\{1,2,3\}$. If $U < \bar{\tau}(x)$, the individual chooses a plan with lower coverage (either plan 2 or 3)

[8] Goeree (2008) also considers the consumers' information heterogeneity. Here, we simplify the specification of ϕ_l by abstracting from it.

depending on the realization of C. Similarly, if $U \geq \bar{\tau}(x)$, the individual chooses either plan 1 or 2 depending on C. Hence, the model's prediction is

$$G(U|X;\theta) = \begin{cases} \{2,3\} & \text{if } U < \bar{\tau}(X) \\ \{1,2\} & \text{if } U \geq \bar{\tau}(X). \end{cases} \tag{3.39}$$

Suppose U has the distribution F_θ. Then, the sharp identifying restrictions are[9]

$$Q(\{2,3\}|x) \geq \nu_\theta(\{2,3\}|x) = F_\theta(G(U|X;\theta) \subseteq \{2,3\}|x) = F_\theta(U < \bar{\tau}(x)) \tag{3.40}$$

$$Q(\{1,2\}|x) \geq \nu_\theta(\{1,2\}|x) = F_\theta(G(U|X;\theta) \subseteq \{1,2\}|x) = F_\theta(U \geq \bar{\tau}(x)). \tag{3.41}$$

Let $\eta(x;\theta) \equiv F_\theta(U < \bar{\tau}(x))$ and $p_{0,j|kl}(j|x) = \frac{p_{0,y|x}(j|x)}{p_{0,y|x}(k|x)+p_{0,y|x}(l|x)}$. Solving the inner optimization problem in (3.16) yields the following profiled-likelihood:

$$q^*_{\theta_2,y|x}(1|x) = p_{0,y|x}(1|x)\mathbb{I}^1(x;\theta) + p_{0,1|12}(1|x)\eta(\theta;x)\mathbb{I}^2(x;\theta) + \eta(\theta;x)\mathbb{I}^3(x;\theta) \tag{3.42}$$

$$q^*_{\theta_2,y|x}(3|x) = p_{0,y|x}(3|x)\mathbb{I}^1(x;\theta) + [1-\eta(\theta;x)]\mathbb{I}^2(x;\theta) + p_{0,3|23}(1|x)[1-\eta(\theta;x)]\mathbb{I}^3(x;\theta), \tag{3.43}$$

where

$$\mathbb{I}^1(x;\theta) \equiv 1\{p_{0,y|x}(1|x) \leq \eta(\theta;x) \leq p_{0,y|x}(1|x) + p_{0,y|x}(2|x)\}$$

$$\mathbb{I}^2(x;\theta) \equiv 1\{\eta(\theta;x) > p_{0,y|x}(1|x) + p_{0,y|x}(2|x)\}$$

$$\mathbb{I}^3(x;\theta) \equiv 1\{p_{0,y|x}(1|x) > \eta(\theta;x)\}.$$

3.5 Monte Carlo Experiments

We conduct Monte Carlo experiments to evaluate the size and power of our cross-fit QLR test. Consider the entry game example. We let $X^{(j)} = (1, X_2^{(j)})'$, where $X_2^{(j)}$

[9] The core determining class for this example is $\{\{2,3\},\{1,2\}\}$ (Barseghyan et al., 2021, Corollary S1.1).

is a player-specific random variable that is either binary or continuously distributed. Similarly, $\beta^{(j)} = (\beta_1^{(j)}, \beta_2^{(j)})'$. The payoff of player j is

$$\pi^{(j)} = Y^{(j)}(\beta_1^{(j)} + \beta_2^{(j)} X_2^{(j)} + \Delta^{(j)} Y^{(-j)} + U^{(j)}),$$

where $(U^{(1)}, U^{(2)}) \sim N(0, I_2)$. In this example, $\theta = (\Delta^{(1)}, \Delta^{(2)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_1^{(2)}, \beta_2^{(2)})'$. We consider two DGPs:

- DGP1: $X_2^{(j)} \overset{i.i.d.}{\sim}$ Bernoulli(0.5).
- DGP2: $X_2^{(j)} \overset{i.i.d.}{\sim} N(0,1)$.

The outcomes are generated by the first model ($s = 1$) with $\theta_0 = (\Delta^{(1)}, \Delta^{(2)}, .5, .5, .5, .5)'$, $\Delta^{(j)} \leq 0, j = 1, 2$. We use a selection mechanism that selects $(1, 0)$ with probability $\tau = 0.5$ whenever the model predicts multiple equilibria. That is, we set $p_{0,y|x}$ to the following:

$$\begin{aligned}
p_{0,y|x}((0,0)|x) &= q_{\theta_0,y|x}((0,0)|x) = [1 - \Phi(x^{(1)\prime}\beta^{(1)})][1 - \Phi(x^{(2)\prime}\beta^{(2)})], \\
p_{0,y|x}((0,1)|x) &= q_{\theta_0,y|x}((0,1)|x) = \eta_1^1(\theta_0; x) - q_{\theta_0,y|x}((1,0)|x), \\
p_{0,y|x}((1,0)|x) &= q_{\theta_0,y|x}((1,0)|x) = \eta_1^3(\theta_0; x) + \tau(\eta_1^2(\theta_0; x) - \eta_1^3(\theta_0; x)), \\
p_{0,y|x}((1,1)|x) &= q_{\theta_0,y|x}((1,1)|x) = \Phi(x^{(1)\prime}\beta^{(1)} + \Delta^{(1)})\Phi(x^{(2)\prime}\beta^{(2)} + \Delta^{(2)}).
\end{aligned}$$

The null hypothesis holds when $\Delta^{(1)} = \Delta^{(2)} = 0$. For local alternatives, we consider a drifting sequence $\Delta^{(1)} = \Delta^{(2)} = -h/\sqrt{n}$ for $h \in \mathbb{N}^+$. For $\hat{p}_{n,y|x}$, in DGP1 we use a cell mean estimator, and in DGP2 we use a sieve Logistic estimator with 3rd-order (tensor product) Hermite polynomials in $(X_2^{(1)}, X_2^{(2)})$ as sieve basis and L^2 penalty. We follow Algorithm 1 to calculate the cross-fit QLR-test statistic $\hat{T}_n$. We set $C = 10$.

For comparison, we consider two alternative tests. The first is a cross-fit QLR

test without regularization:

$$\tilde{T}_n = \frac{\tilde{T}_{I_1} + \tilde{T}_{I_2}}{\sqrt{2}}, \text{ where } \tilde{T}_{I_\ell} = \frac{\sqrt{n/2}\widehat{\mathrm{QLR}}_{I_\ell}}{\hat{\sigma}_{I_\ell}(\hat{\theta}_{I_{-\ell}})} \text{ for } \ell = 1, 2.$$

The second is the test proposed by Hsu and Shi (2017) (henceforth HS). Their test statistic is given by

$$\hat{T}_n^{\mathrm{HS}} = \frac{\sqrt{n}\widehat{\mathrm{LR}}_n + \hat{\omega}_n U}{\sqrt{\hat{\sigma}_n^2 + \hat{\omega}_n^2}},$$

where $\widehat{\mathrm{LR}}_n$ is the sample analog of the difference between the average generalized empirical likelihood (AGEL) distances from the two models to the true DGP, and $U \sim N(0,1)$ is independent of the original sample. Their benchmark data-dependent choice of $\hat{\omega}_n$ is

$$\hat{\omega}_n = (2 + 2 \cdot C^{2d_X} t_{b_n}^{-2} \hat{\sigma}_n^2)^{-1}$$

with $b_n = n/\log(n)$, $t_n = n^{-1/(4d_X+2)}$, and $C = 5$. Since their test is designed for models defined by conditional moment restrictions with continuous conditioning variables, we focus on DGP2.

We consider sample sizes $n = 1000, 500, 250$. We focus on two-sided tests and calculate rejection probabilities based on 5000 Monte Carlo repetitions. Tables 3.1–3.3 and Figure 3·3 report the results. We observe that for each n, our cross-fit QLR test has the correct size while the cross-fit QLR test without regularization severely underrejects. The HS test tends to overreject when the sample size is small ($n = 250$), although the size distortion appears to diminish as the sample size increases. As h grows, our cross-fit QLR test has nontrivial power, almost matching that of the HS test, albeit less than the cross-fit QLR test without regularization. This power discrepancy becomes more pronounced for larger values of n. Table 3.4 reports the average runtime for computing our cross-fit QLR test and the HS test across different

values of h and 5000 Monte Carlo repetitions.[10] On average, the HS test takes about 160 times longer than ours. The computational burden of the HS test arises from the duality result underlying $\widehat{\mathrm{LR}}_n$, which necessitates two nested optimization loops over both the model parameter and the Lagrange multiplier. Overall, our test has advantages in terms of small-sample performance and computational costs.

3.6 Concluding remarks

This chapter expands the scope of likelihood-based model selection tests to incomplete models. A novel feature is the use of the profiled likelihood that allows the researcher to compare parametric discrete choice models regardless of their model completeness or incompleteness. The proposed QLR statistic is asymptotically normally distributed and provides a tractable, uniformly valid test to select a parametric model. A Monte Carlo experiment demonstrates that the proposed test performs well in controlling its size, has competitive power, and offers computational advantages compared to existing methods.

[10] The simulations use the replication code of Hsu and Shi (2017) with minimal adaption and are run on the Boston University Shared Computing Cluster (SCC) with 28 cores.

Table 3.1: Rejection Probabilities ($n = 1000$)

Tests	Size	Power (values of $-h/\sqrt{n}$ below)									
		-0.032	-0.063	-0.095	-0.126	-0.158	-0.19	-0.221	-0.253	-0.285	-0.316
Panel A: DGP1 (Discrete X)											
$\hat{T}_n$	0.045	0.045	0.046	0.048	0.048	0.053	0.062	0.094	0.151	0.259	0.397
$\tilde{T}_n$	0.001	0.001	0.007	0.030	0.088	0.231	0.415	0.644	0.833	0.942	0.983
Panel B: DGP2 (Continuous X)											
$\hat{T}_n$	0.046	0.045	0.045	0.044	0.049	0.058	0.082	0.129	0.218	0.350	0.535
$\tilde{T}_n$	0.000	0.002	0.008	0.028	0.109	0.246	0.477	0.700	0.864	0.952	0.989
$\hat{T}_n^{\text{HS}}$	0.054	0.053	0.054	0.054	0.056	0.062	0.078	0.123	0.203	0.346	0.530

Table 3.2: Rejection Probabilities ($n = 500$)

Tests	Size	Power (values of $-h/\sqrt{n}$ below)									
		-0.045	-0.089	-0.134	-0.179	-0.224	-0.268	-0.313	-0.358	-0.402	-0.447
Panel A: DGP1 (Discrete X)											
$\hat{T}_n$	0.044	0.047	0.044	0.048	0.051	0.068	0.113	0.203	0.337	0.538	0.711
$\tilde{T}_n$	0.001	0.003	0.011	0.039	0.108	0.254	0.465	0.675	0.834	0.941	0.982
Panel B: DGP2 (Continuous X)											
$\hat{T}_n$	0.045	0.047	0.045	0.048	0.056	0.086	0.154	0.267	0.453	0.642	0.801
$\tilde{T}_n$	0.001	0.001	0.009	0.036	0.121	0.263	0.488	0.705	0.871	0.950	0.983
$\hat{T}_n^{\text{HS}}$	0.057	0.056	0.058	0.061	0.066	0.088	0.147	0.262	0.430	0.609	0.751

Table 3.3: Rejection Probabilities ($n = 250$)

Tests	Size	Power (values of $-h/\sqrt{n}$ below)									
		-0.063	-0.126	-0.19	-0.253	-0.316	-0.379	-0.442	-0.506	-0.569	-0.632
Panel A: DGP1 (Discrete X)											
$\hat{T}_n$	0.049	0.048	0.050	0.047	0.066	0.109	0.208	0.375	0.576	0.302	0.408
$\tilde{T}_n$	0.002	0.006	0.018	0.050	0.138	0.274	0.473	0.668	0.834	0.401	0.462
Panel B: DGP2 (Continuous X)											
$\hat{T}_n$	0.045	0.046	0.049	0.050	0.089	0.153	0.284	0.456	0.654	0.814	0.920
$\tilde{T}_n$	0.002	0.003	0.015	0.046	0.148	0.296	0.505	0.700	0.852	0.940	0.980
$\hat{T}_n^{\text{HS}}$	0.090	0.090	0.092	0.095	0.120	0.177	0.289	0.446	0.610	0.765	0.882

Table 3.4: Average Runtime (in sec.)

	$\hat{T}_n$	$\hat{T}_n^{\text{HS}}$
$n = 1000$	0.42	58.60
$n = 500$	0.39	69.28
$n = 250$	0.48	79.22

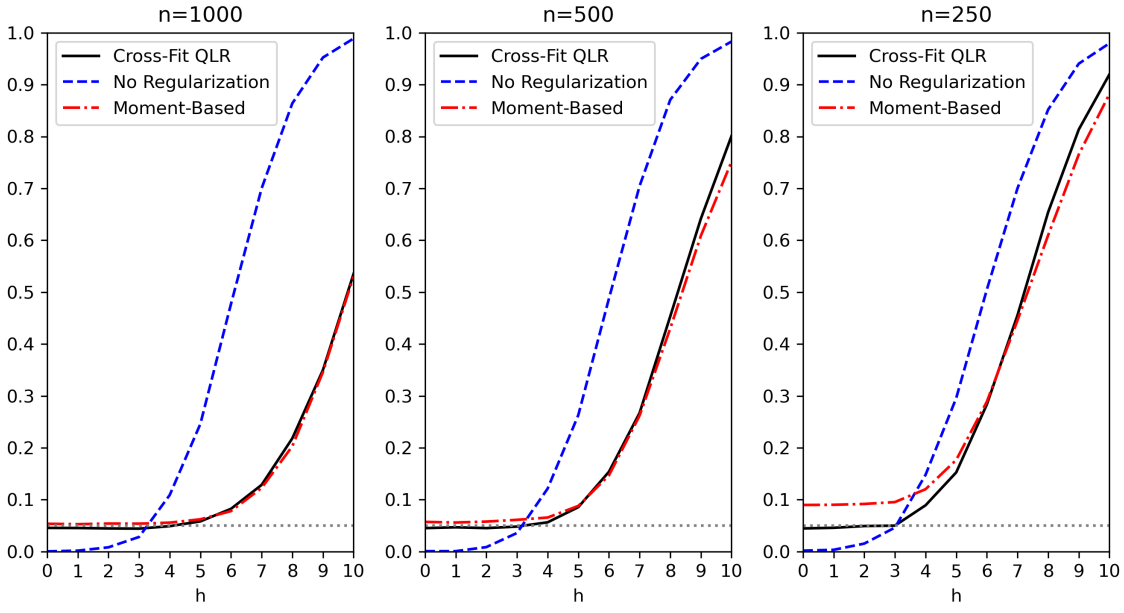


Figure 3·3: Power Curves

Appendix A

Appendix for Chapter 1

A.1 Proofs

Proof of Theorem 1.1. Write $U = (U_1, \ldots, U_T)$. Following Chesher and Rosen (2017), I adopt the notion of structures. In my case, a structure is a pair $(\theta, \mathcal{F}_{U|X})$. Let $\mathcal{M}$ be the set of structures that satisfy Assumptions 1.1 and 1.2. 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, $(\theta, \mathcal{F}_{U|X}) \in \mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ if $(\theta, \mathcal{F}_{U|X})$ is admitted by $\mathcal{M}$ and $\mathcal{F}_{Y|X}$ can be reproduced by $(\theta, \mathcal{F}_{U|X})$. Then, the sharp identified set for $\mathcal{F}_{Y_t(\underline{x})|X}$ is defined as

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

Taking $\mathcal{F}_{Y_t(\underline{x})|X}$ from the right-hand side of (1.1), I want to show that $\mathcal{F}_{Y_t(\underline{x})|X} \in \mathsf{F}^*_{Y_t(\underline{x})|X}$, which amounts to exhibiting $(\theta, \mathcal{F}_{U|X}) \in \mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ satisfying

$$\forall \mathcal{T} \in \mathsf{F}(\mathcal{Y}), F_{Y_t(\underline{x})|X=x}(\mathcal{T}) = F_{U_t|X=x}(\mathcal{U}(\mathcal{T}, \underline{x}; \theta)) \text{ a.e. } x \in \text{Supp}(X). \tag{A.1}$$

By Assumption 1.2, the projection of $\mathcal{I}(\mathcal{M}, \mathcal{F}_{Y|X})$ on Θ is $\{\theta_0\}$. By (1.1), there exists $\mathcal{F}_{U_t|X} \in \mathsf{F}^*_{U_t|X}$ such that (A.1) holds. Let $\mathcal{F}_{U_{t'}|X} = \mathcal{F}_{U_t|X}$ $\forall t' \in \{1, \ldots, T\}$. Fixing any $x \in \text{Supp}(X)$, by the definition of $\mathsf{F}^*_{U_t|X}$, θ_0 and $\{F_{U_{t'}|X=x}\}_{t'=1}^T$ reproduces the marginals of $F_{Y|X=x}$, while by Sklar's theorem, there exists a T-variate copula $C_X(\cdot|x)$ that reproduces the dependence structure of $F_{Y|X=x}$, i.e.,

$$F_{Y|X=x} = C_X(F_{Y_1|X=x}, \ldots, F_{Y_T|X=x}|x).$$

Let $\mathcal{C}_X = \{C_X(\cdot|x) : x \in \text{Supp}(X)\}$. It remains to construct $\mathcal{F}_{U|X}$ from $\{\mathcal{F}_{U_{t'}|X}\}_{t'=1}^T$ and $\mathcal{C}_X$. For any $x \in \text{Supp}(X)$, define

$$\begin{aligned} \bar{g}_{x_{t'},\theta}(u_{t'}) &= g(x_{t'}, u_{t'}; \theta),\ t' \in \{1, \ldots, T\}, \\ G(x, u; \theta) &= (g(x_1, u_1; \theta), \ldots, g(x_T, u_T; \theta)), \\ \bar{G}_{x,\theta}(u) &= G(x, u; \theta). \end{aligned}$$

Fix $x \in \text{Supp}(X)$. Then, $F_{Y|X=x}$ is the pushforward of $F_{U|X=x}$ by $\bar{G}_{x,\theta_0}$, and for each t', $F_{Y_{t'}|X=x}$ is the pushforward of $F_{U_{t'}|X=x}$ by $\bar{g}_{x_{t'},\theta_0}$; write

$$\begin{aligned} F_{Y|X=x} &= (\bar{G}_{x,\theta_0})_{\#} F_{U|X=x}, \\ F_{Y_{t'}|X=x} &= (\bar{g}_{x_{t'},\theta_0})_{\#} F_{U_{t'}|X=x},\ t' \in \{1, \ldots, T\}. \end{aligned}$$

Now, one can pick $F_{U|X=x}$ such that

$$(\bar{G}_{x,\theta_0})_{\#} F_{U|X=x} = C_X((\bar{g}_{x_1,\theta_0})_{\#} F_{U_1|X=x}, \ldots, (\bar{g}_{x_T,\theta_0})_{\#} F_{U_T|X=x}|x).$$

□

Proof of (1.5). 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 1.2. By definition, $\mathcal{F}_{U_t|X} \in \mathsf{F}^*_{U_t|X}$ if and only if $\forall y' \in \mathcal{Y}$, $\forall t' \in \{1, \ldots, T\}$,

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

It follows that

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

where the second equality follows from (1.3). Taking $\mathcal{F}_{Y_t(\underline{x})|X}$ from the right-hand side of (1.6), I want to show that $\mathcal{F}_{Y_t(\underline{x})|X} \in \mathsf{F}^*_{Y_t(\underline{x})|X}$, which amounts to exhibiting $F_{U_t|X=x}$ for all $x \in \text{Supp}(X)$ satisfying $\forall y \in \mathcal{Y}$,

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

Fix $x \in \text{Supp}(X)$. 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, (1.6) 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 1.2 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 $t', t'' \in \{1, \ldots, T+1\}$ and $y, y' \in \mathcal{Y}$,

$$\underline{u}_{t'}(y) \leq \underline{u}_{t''}(y') \iff p_{t'}(y) \geq p_{t''}(y'). \tag{A.2}$$

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

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

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 (A.2). □

Proof of Theorem 1.3. Taking $\mathcal{F}_{Y_t(\underline{x})|X}$ from the right-hand side of (1.7), I want to show that $\mathcal{F}_{Y_t(\underline{x})|X} \in \mathsf{F}^*_{Y_t(\underline{x})|X}$, which amounts to exhibiting $F_{U_t|X=x}$ for all $x \in \text{Supp}(X)$ 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

$$\sum_{(j_1,\ldots,j_T,j'):\ \mathcal{U}_{j_1,\ldots,j_T,j'} \neq \emptyset,\ j'=j} q_{j_1,\ldots,j_T,j'} = F_{Y_t(\underline{x})|X=x}(\{j\}), \tag{A.3}$$

$$\sum_{(j_1,\ldots,j_T,j'):\ \mathcal{U}_{j_1,\ldots,j_T,j'} \neq \emptyset,\ j_{t'}=j} q_{j_1,\ldots,j_T,j'} = F_{Y_{t'}|X=x}(\{j\}), \tag{A.4}$$

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^{ct} be the matrix with elements in $\{0, 1\}$ such that (A.3) 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 (A.4) 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 $\{z^{t'} = (z^{t'}_0, z^{t'}_1, \ldots, z^{t'}_J)^\top\}_{t'=1}^T$ and $w = (w_0, w_1, \ldots, w_J)^\top$ be $(J+1)$-dimensional constant vectors. Farkas's Lemma states that if

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

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 $M^{-1}_{t'}$ denote the inverse mapping. Then, $(\{M^{-1}_{t'}(J), \ldots, M^{-1}_{t'}(j)\}$,

$\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}) \in \mathbb{Y}(\underline{x}^\top \beta_0, x_{t'}^\top \beta_0)$ for $j > 0$. For any $\{a_j^{t'}\}_{j=0,1,\ldots,J,t'=1,\ldots,T} \in \mathbb{R}$,

$$
\begin{aligned}
& w^\top p^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\top p_{t'}^{\mathrm{ob}} \\
= \; & \sum_{j=0}^{J} w_j F_{Y_t(\underline{x})|X=x}(\{j\}) + \sum_{t'=1}^{T} \sum_{j=0}^{J} z_j^{t'} F_{Y_{t'}|X=x}(\{j\}) \\
= \; & \sum_{t'=1}^{T} \sum_{j=0}^{J} a^{t'}_{M_{t'}^{-1}(j)} \\
& \cdot \underbrace{\left(F_{Y_{t'}|X=x}(\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\}) - F_{Y_t(\underline{x})|X=x}(\{M_{t'}^{-1}(J), \ldots, M_{t'}^{-1}(j)\})\right)}_{\geq 0 \text{ by } (1.7)} \\
& + \sum_{j=0}^{J} \left(w_j + \sum_{t'=1}^{T} \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \right) F_{Y_t(\underline{x})|X=x}(\{j\}) \\
& + \sum_{t'=1}^{T} \sum_{j=0}^{J} \left(z_j^{t'} - \sum_{\ell: M_{t'}(\ell) \leq M_{t'}(j)} a_\ell^{t'} \right) F_{Y_{t'}|X=x}(\{j\}).
\end{aligned}
$$

Therefore, given $w^\top Q^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\top Q_{t'}^{\mathrm{ob}} \geq 0$, $w^\top p^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\top p_{t'}^{\mathrm{ob}} \geq 0$ 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^{ct} and $Q_1^{\mathrm{ob}}, \ldots, Q_T^{\mathrm{ob}}$, $w^\top Q^{\mathrm{ct}} + \sum_{t'=1}^{T} (z^{t'})^\top Q_{t'}^{\mathrm{ob}} \geq 0$ yields

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

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

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

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

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

Now it can be concluded that (1.7) holds. □

Proof of Theorem 1.4. 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

$$\begin{aligned}
&\sqrt{N}(\hat{\Psi}(\theta) - \Psi(\theta)) \\
&= \mathbb{G}_N\Big(\sum_{\lambda \in \Lambda(X;\theta)} 1\{\lambda^*(X;\theta,\tau_0) = \lambda\} \lambda^\top I(Y) \Big) && \text{(A.5)} \\
&\quad + \mathbb{G}_N\Big(\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\hat{\tau}) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\}) \lambda^\top I(Y) \Big) && \text{(A.6)} \\
&\quad + \sqrt{N} E\Big[\sum_{\lambda \in \Lambda(X;\theta)} (1\{\lambda^*(X;\theta,\hat{\tau}) = \lambda\} - 1\{\lambda^*(X;\theta,\tau_0) = \lambda\}) \lambda^\top I(Y) \Big]. && \text{(A.7)}
\end{aligned}$$

To show (A.6) and (A.7) are $o_p(1)$, I will use the following lemma:

Lemma A.1. *Suppose that Assumptions 1.3 and 1.5 hold. Then, for all θ, there*

exists $C > 0$ *such that for any* $\delta \geq 0$,

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

First, by Assumption 1.6, (A.6) is $o_p(1)$ if the stochastic equicontinuity property holds: for all positive values $\delta_N = o(1)$,

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

To this end, note that by Assumption 1.3,

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

where

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

It follows that

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

By Lemma A.1 and Theorem 3 of Chen et al. (2003), (A.6) is $o_p(1)$. Second, for

(A.7), observe that

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

where the last inequality follows from Lemma A.1. Then, by Assumption 1.6, (A.7) is $o_p(1)$. Now I can apply the central limit theorem to (A.5) to obtain the desired result. □

A.2 Monte Carlo Simulation

I consider the same data generating process as in Section 1.6 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 procedure proposed by Cattaneo et al. (2020) to construct the bootstrap version for $\beta_0^{(2)}$.[1] I choose $N = 5000$ and $T = 10$, with $S = 1000$ simulations and $B = 399$ bootstrap replications. I set the nominal level $\alpha = 0.05$.

[1] Due to computational challenges, I plug in the true Hessian matrix rather than the estimated one. I leave for future work how the proposed confidence interval performs when the sampling uncertainty is also present in the Hessian matrix.

In Table A.1, 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 (1.9). For comparison, I also consider an infeasible scenario where the true value of $\tau_0(x)$ is known. One can see that the coverage rates are close to 95% when $N = 5000$ but tend to drop when N is reduced to 1000, which seems to be driven by the undercoverage of $\beta_0^{(2)}$. The estimation error of τ_0 has minimal effects on both coverage and confidence interval length.

Table A.1: 95% CI for Sharp Bounds on $\Pr(Y_{it}(\underline{x}) = 1)$

	Coverage	Avg. Excess Length	Coverage for $\beta_0^{(2)}$
Panel A: $N = 5000$			
true CCP	0.948	0.049	0.952
estimated CCP	0.949	0.049	
Panel B: $N = 1000$			
true CCP	0.920	0.091	0.901
estimated CCP	0.918	0.091	

Appendix B

Appendix for Chapter 2

B.1 Proofs

For a binary encouragement rule π described in Section 2.2.3, define an ideal version of the empirical welfare criterion and budget based on the true propensity score and MTE as

$$\bar{W}_n(\pi) = E_n\Big[\pi(X,Z)\int_{p(X,\alpha_0(Z))}^{p(X,\alpha_1(Z))} \text{MTE}(u,X)\,du\Big],$$

$$B_n(\pi) = E_n[C(X,Z)\cdot\{\pi(X,Z)\cdot p(X,\alpha_1(Z)) + (1-\pi(X,Z))\cdot p(X,\alpha_0(Z))\}].$$

Proof of Theorem 2.1. By (2.2), we can write

$$\begin{aligned}
Y(D(\boldsymbol{\alpha}(X,Z))) &= Y(1)\cdot D(\boldsymbol{\alpha}(X,Z)) + Y(0)\cdot(1-D(\boldsymbol{\alpha}(X,Z))) \\
&= Y(1)\cdot 1\{p(X,\boldsymbol{\alpha}(X,Z)) \geq U\} + Y(0)\cdot 1\{p(X,\boldsymbol{\alpha}(X,Z)) < U\} \\
&= Y(1)\cdot 1\{p(X,Z) \geq U\} + Y(0)\cdot 1\{p(X,Z) < U\} \\
&\quad + (Y(1)-Y(0))\cdot(1\{p(X,\boldsymbol{\alpha}(X,Z)) \geq U\} - 1\{p(X,Z)\geq U\}) \\
&= Y + (Y(1)-Y(0))\cdot(1\{p(X,\boldsymbol{\alpha}(X,Z)) \geq U\} - 1\{p(X,Z)\geq U\}).
\end{aligned}$$

Hence,

$$\begin{aligned}
W(\boldsymbol{\alpha}) - E[Y] &= E[(Y(1) - Y(0)) \cdot (1\{p(X, \boldsymbol{\alpha}(X, Z)) \geq U\} - 1\{p(X, Z) \geq U\})] \\
&= E[E[E[Y(1) - Y(0)|X, Z, U] \\
&\quad \cdot (1\{p(X, \boldsymbol{\alpha}(X, Z)) \geq U\} - 1\{p(X, Z) \geq U\})|X, Z]] \\
&= E[\mathrm{MTE}(U, X) \cdot E[1\{p(X, \boldsymbol{\alpha}(X, Z)) \geq U\} - 1\{p(X, Z) \geq U\}|X, Z]] \\
&= E\Big[\int_0^1 \mathrm{MTE}(u, X) \cdot (1\{p(X, \boldsymbol{\alpha}(X, Z)) \geq u\} - 1\{p(X, Z) \geq u\})\, du\Big],
\end{aligned}$$

where the second equality follows from the law of iterated expectations, the third equality follows from Assumption 2.1(ii) and the definition of the MTE, and the fourth equality follows from $U|X, Z \sim \mathrm{Unif}[0, 1]$. □

Proof of Theorem 2.2. We have for any $\pi' \in \Pi$,

$$\begin{aligned}
W(\pi') - W(\hat{\pi}_{\mathrm{FEWM}}) &= \bar{W}(\pi') - \bar{W}(\hat{\pi}_{\mathrm{FEWM}}) \\
&= [\bar{W}(\pi') - \bar{W}_n(\pi')] + [\bar{W}_n(\pi') - \hat{W}_n(\pi')] \\
&\quad + [\hat{W}_n(\pi') - \hat{W}_n(\hat{\pi}_{\mathrm{FEWM}})] \\
&\quad + [\hat{W}_n(\hat{\pi}_{\mathrm{FEWM}}) - \bar{W}_n(\hat{\pi}_{\mathrm{FEWM}})] + [\bar{W}_n(\hat{\pi}_{\mathrm{FEWM}}) - \bar{W}(\hat{\pi}_{\mathrm{FEWM}})] \\
&\leq [\bar{W}(\pi') - \bar{W}_n(\pi')] + [\bar{W}_n(\pi') - \hat{W}_n(\pi')] \\
&\quad + [\hat{W}_n(\hat{\pi}_{\mathrm{FEWM}}) - \bar{W}_n(\hat{\pi}_{\mathrm{FEWM}})] + [\bar{W}_n(\hat{\pi}_{\mathrm{FEWM}}) - \bar{W}(\hat{\pi}_{\mathrm{FEWM}})] \\
&\leq 2 \sup_{\pi \in \Pi} |\bar{W}_n(\pi) - \bar{W}(\pi)| + 2 \sup_{\pi \in \Pi} |\hat{W}_n(\pi) - \bar{W}_n(\pi)|,
\end{aligned}$$

where the first inequality follows from $\hat{W}_n(\hat{\pi}_{\mathrm{FEWM}}) \geq \hat{W}_n(\pi')$. Hence,

$$E[R(\hat{\pi}_{\mathrm{FEWM}})] \leq 2E\Big[\sup_{\pi \in \Pi} |\bar{W}_n(\pi) - \bar{W}(\pi)|\Big] + 2E\Big[\sup_{\pi \in \Pi} |\hat{W}_n(\pi) - \bar{W}_n(\pi)|\Big].$$

First, we bound $E[\sup_{\pi \in \Pi} |\bar{W}_n(\pi) - \bar{W}(\pi)|]$. Define

$$\mathcal{F} = \Big\{ f_\pi(x, z) = \pi(x, z) \int_{p(x, \alpha_0(z))}^{p(x, \alpha_1(z))} \mathrm{MTE}(u, x)\, du : \pi \in \Pi \Big\}.$$

By Assumption 2.3 and Lemma A.1 of Kitagawa and Tetenov (2018a), $\mathcal{F}$ is a VC-subgraph class of uniformly bounded functions with VC-dimension less than or equal

to VC(Π). Then, we can apply Lemma A.4 of Kitagawa and Tetenov (2018a) to obtain

$$E\Big[\sup_{\pi\in\Pi}|\bar{W}_n(\pi)-\bar{W}(\pi)|\Big]=E\Big[\sup_{f\in\mathcal{F}}|E_n[f(X,Z)]-E[f(X,Z)]|\Big]\leq C_1\bar{M}\sqrt{\frac{\mathrm{VC}(\Pi)}{n}}, \tag{B.1}$$

where C_1 is a universal constant. Next, we bound $E[\sup_{\pi\in\Pi}|\hat{W}_n(\pi)-\bar{W}_n(\pi)|]$. By the triangle inequality, for any $\pi\in\Pi$,

$$\begin{aligned}|\hat{W}_n(\pi)-\bar{W}_n(\pi)|\leq{}& E_n\Big[\Big|\int_{p(X,\alpha_0(Z))}^{p(X,\alpha_1(Z))}(\widehat{\mathrm{MTE}}(u,X)-\mathrm{MTE}(u,X))\,du\Big|\Big]\\&+\sum_{d\in\{0,1\}}E_n\Big[\Big|\int_{p(X,\alpha_d(Z))}^{\hat{p}(X,\alpha_d(Z))}\widehat{\mathrm{MTE}}(u,X)\,du\Big|\Big].\end{aligned} \tag{B.2}$$

The first term on the right-hand side of (B.2) can be bounded as

$$E_n\Big[\Big|\int_{p(X,\alpha_0(Z))}^{p(X,\alpha_1(Z))}(\widehat{\mathrm{MTE}}(u,X)-\mathrm{MTE}(u,X))\,du\Big|\Big]\leq E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)-\mathrm{MTE}(u,X)|\Big].$$

Each summand in the second term on the right-hand side of (B.2) can be bounded as

$$\begin{aligned}E_n\Big[\Big|\int_{p(X,\alpha_d(Z))}^{\hat{p}(X,\alpha_d(Z))}\widehat{\mathrm{MTE}}(u,X)\,du\Big|\Big]&\leq E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)||\hat{p}(X,\alpha_d(Z))-p(X,\alpha_d(Z))|\Big]\\&\leq E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)|^2\Big]^{1/2}\\&\quad\cdot E_n[|\hat{p}(X,\alpha_d(Z))-p(X,\alpha_d(Z))|^2]^{1/2},\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality. Since these bounds do not depend on π, we can apply the Cauchy-Schwarz inequality again to

obtain

$$
\begin{aligned}
E\Big[\sup_{\pi\in\Pi}|\hat{W}_n(\pi)-\bar{W}_n(\pi)|\Big] &\le E\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)-\mathrm{MTE}(u,X)|\Big]\Big] \\
&\quad + E\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)|^2\Big]\Big]^{1/2} \\
&\quad \cdot \sum_{d\in\{0,1\}} E[E_n[|\hat{p}(X,\alpha_d(Z))-p(X,\alpha_d(Z))|^2]]^{1/2} \\
&= O(\psi_n^{-1}\vee\phi_n^{-1}), \qquad \text{(B.3)}
\end{aligned}
$$

where the equality follows from Assumption 2.4. □

Proof of Theorem 2.3. For any $\epsilon>0$,

$$
\begin{aligned}
\Pr(W(\pi_{\mathrm{B}}^*)-W(\hat{\pi}_{\mathrm{BEWM}})>\epsilon) &= \Pr(\bar{W}(\pi_{\mathrm{B}}^*)-\bar{W}(\hat{\pi}_{\mathrm{BEWM}})>\epsilon,\hat{B}_n(\pi_{\mathrm{B}}^*)\le\kappa) \\
&\quad + \Pr(\bar{W}(\pi_{\mathrm{B}}^*)-\bar{W}(\hat{\pi}_{\mathrm{BEWM}})>\epsilon,\hat{B}_n(\pi_{\mathrm{B}}^*)>\kappa). \qquad \text{(B.4)}
\end{aligned}
$$

Noting that $\hat{B}_n(\pi_{\mathrm{B}}^*)\le\kappa$ implies $\hat{W}_n(\hat{\pi}_{\mathrm{BEWM}})\ge\hat{W}_n(\pi_{\mathrm{B}}^*)$, the first term on the right-hand side of (B.4) can be bounded as

$$
\begin{aligned}
&\Pr(\bar{W}(\pi_{\mathrm{B}}^*)-\bar{W}(\hat{\pi}_{\mathrm{BEWM}})>\epsilon,\hat{B}_n(\pi_{\mathrm{B}}^*)\le\kappa) \\
&\le \Pr(\bar{W}(\pi_{\mathrm{B}}^*)-\hat{W}_n(\pi_{\mathrm{B}}^*)+\hat{W}_n(\hat{\pi}_{\mathrm{BEWM}})-\bar{W}(\hat{\pi}_{\mathrm{BEWM}})>\epsilon,\hat{B}_n(\pi_{\mathrm{B}}^*)\le\kappa) \\
&\le \Pr\Big(2\sup_{\pi\in\Pi}|\bar{W}(\pi)-\hat{W}_n(\pi)|>\epsilon\Big). \qquad \text{(B.5)}
\end{aligned}
$$

The second term on the right-hand side of (B.4) can be bounded as

$$
\begin{aligned}
\Pr(\bar{W}(\pi_{\mathrm{B}}^*)-\bar{W}(\hat{\pi}_{\mathrm{BEWM}})>\epsilon,\hat{B}_n(\pi_{\mathrm{B}}^*)>\kappa) &\le \Pr(\hat{B}_n(\pi_{\mathrm{B}}^*)>\kappa) \\
&\le \Pr(\hat{B}_n(\pi_{\mathrm{B}}^*)>B(\pi_{\mathrm{B}}^*)) \\
&\le \sup_{\epsilon'>0}\Pr(\hat{B}_n(\pi_{\mathrm{B}}^*)-B(\pi_{\mathrm{B}}^*)>\epsilon') \\
&\le \sup_{\epsilon'>0}\Pr\Big(\sup_{\pi\in\Pi}|B(\pi)-\hat{B}_n(\pi)|>\epsilon'\Big), \qquad \text{(B.6)}
\end{aligned}
$$

where the second inequality follows from $B(\pi_{\mathrm{B}}^*) \leq \kappa$. Also, for any $\epsilon > 0$,

$$\begin{aligned}\Pr(B(\hat{\pi}_{\mathrm{BEWM}}) - \kappa > \epsilon) &\leq \Pr(B(\hat{\pi}_{\mathrm{BEWM}}) - \hat{B}_n(\hat{\pi}_{\mathrm{BEWM}}) > \epsilon) \\ &\leq \Pr\Big(\sup_{\pi\in\Pi} |B(\pi) - \hat{B}_n(\pi)| > \epsilon\Big), \end{aligned} \tag{B.7}$$

where the first inequality follows from $\hat{B}_n(\hat{\pi}_{\mathrm{BEWM}}) \leq \kappa$. By the triangle inequality, for any $\pi \in \Pi$,

$$|B(\pi) - \hat{B}_n(\pi)| \leq |B(\pi) - B_n(\pi)| + |B_n(\pi) - \hat{B}_n(\pi)|. \tag{B.8}$$

For the first term on the right-hand side of (B.8), from the same argument as the proof of (B.1), we can apply Lemma A.4 of Kitagawa and Tetenov (2018a) to obtain

$$E\Big[\sup_{\pi\in\Pi} |B_n(\pi) - B(\pi)|\Big] = O(n^{-1/2}). \tag{B.9}$$

For the second term on the right-hand side of (B.8), note that

$$\begin{aligned}\sup_{\pi\in\Pi} |B_n(\pi) - \hat{B}_n(\pi)| \leq E_n[&C(X,Z) \\ &\cdot (|\hat{p}(X,\alpha_1(Z)) - p(X,\alpha_1(Z))| + |\hat{p}(X,\alpha_0(Z)) - p(X,\alpha_0(Z))|)] \\ \leq \sup_{x,z} C(x,z) &\sum_{d\in\{0,1\}} E_n[|\hat{p}(X,\alpha_d(Z)) - p(X,\alpha_d(Z))|].\end{aligned}$$

Hence, by Assumption 2.4(ii),

$$E\Big[\sup_{\pi\in\Pi} |B_n(\pi) - \hat{B}_n(\pi)|\Big] = O(\psi_n^{-1}). \tag{B.10}$$

Combining (B.1), (B.3)–(B.10), and Markov's inequality gives us the desired result. □

B.2 Verification of Assumption 2.2

Verification of Assumption 2.2 in Example 2.1. Assumption 2.2(i) is satisfied because $p(x,z)$ is point-identified over $\mathrm{Supp}(X,Z)$. Assumption 2.2(ii) is satisfied because for every $x \in \mathcal{X}$, $\mathrm{MTE}(\cdot,x)$ is point-identified over $\mathcal{P}(x)$, and $[\min \mathcal{P}_{\boldsymbol{\alpha}}(x), \max \mathcal{P}_{\boldsymbol{\alpha}}(x)] \subset [\min \mathcal{P}(x), \max \mathcal{P}(x)] \subset \mathcal{P}(x)$, where $\mathcal{P}(x)$ denotes the support of $p(X,Z)$ conditional

on $X = x$. □

Verification of Assumption 2.2 in Example 2.2. Assumption 2.2(i) is satisfied because γ is point-identified and $\theta(z)$ is point-identified over $\mathcal{Z}$, provided that $E[(X - E[X|Z])(X - E[X|Z])^\top]$ is positive definite. Under (E2.2-4) and (E2.2-5), we can write the conditional mean of Y given $(X, p(X,Z))$ as

$$E[Y|X = x, p(X,Z) = u] = uX^\top\beta_1 + (1-u)X^\top\beta_0 + E[V|p(X,Z) = u].$$

Provided that $E[(X - E[X|p(X,Z)])(X - E[X|p(X,Z)])^\top]$ is positive definite, β_1 and β_0 are point-identified, and $E[V|p(X,Z) = u]$ is point-identified over $\text{Supp}(p(X,Z))$. Then, Assumption 2.2(ii) is satisfied because for any $(x,u) \in \mathcal{X} \times \text{Supp}(p(X,Z))$, the MTE is point-identified as

$$\text{MTE}(u,x) = x^\top(\beta_1 - \beta_0) + \frac{\partial}{\partial u}E[V|p(X,Z) = u],$$

and (E2.2-1) and (E2.2-2) imply that for every $x \in \mathcal{X}$,

$$\begin{aligned}[\min \mathcal{P}_{\boldsymbol{\alpha}}(x), \max \mathcal{P}_{\boldsymbol{\alpha}}(x)] &\subset \cup_{x\in\mathcal{X}}[\min \mathcal{P}_{\boldsymbol{\alpha}}(x), \max \mathcal{P}_{\boldsymbol{\alpha}}(x)] \\ &\subset \cup_{x\in\mathcal{X}}[\min \mathcal{P}(x), \max \mathcal{P}(x)] \\ &\subset \text{Supp}(p(X,Z)).\end{aligned}$$

□

Verification of Assumption 2.2 in Example 2.3. Assumption 2.2(i) is satisfied because γ is point-identified provided that there is no multicollinearity in $\mu(X,Z)$ elements. Assumption 2.2(ii) is satisfied because

$$\text{MTE}(u,x) = x^\top(\beta_1 - \beta_0) + \sum_{j=2}^{J} j\eta_j u^{j-1},$$

and β_1, β_0, and $\eta_2, \ldots, \eta_J$ are point-identified provided that there is no multicollinearity in $(p(X,Z)X^\top, (1-p(X,Z))X^\top, p(X,Z)^2, \ldots, p(X,Z)^J)$. □

B.3 Encouragement Rules with a Binary Instrument

In many applications, the instrument is binary by construction. For completeness, we present a discussion of cases in which a binary instrument is intervened upon. An encouragement rule is a mapping $\pi : \mathcal{X} \to \{0, 1\}$ that determines the manipulated value of the instrument. For example, the encouragement could be eligibility for welfare programs such as the National Job Training Partnership Act (JTPA) or the 401(k) retirement program. Define the social welfare criterion as $W(\pi) = E[Y(D(\pi(X)))]$. Given a feasible class Π of encouragement rules, the optimal encouragement rule solves $\max_{\pi\in\Pi} W(\pi)$. Under (2.2) and Assumption 2.1, $W(\pi)$ is identified as

$$W(\pi) = E[E[Y|X, Z = 1] \cdot \pi(X) + E[Y|X, Z = 0] \cdot (1 - \pi(X))].$$

Therefore, the optimal encouragement rule is identified without observing D. Intuitively, the optimization problem amounts to finding the optimal treatment rule that assigns Z rather than D, from an intention-to-treat perspective. In this case, Assumption 2.1(i) warrants unconfoundedness, and the analysis essentially follows the original EWM framework.

To understand which subpopulation benefits from the encouragement rule, we impose further assumptions on the selection behavior:

Assumption B.1 (Increasing Propensity Score)**.** *For each $x \in \mathcal{X}$, $p(x, 1) \geq p(x, 0)$.*

Under (2.2) and Assumption B.1, we can partition the population into three compliance groups: always-takers, never-takers, and compliers. Let $\varrho_c(x) = \Pr(D(1) > D(0)|X = x)$ be the conditional population share of compliers and $\Delta_c(x) = E[Y(1) - Y(0)|X = x, D(1) > D(0)]$ be the CATE for compliers. It is worth noting that $W(\pi)$

depends on π only through the counterfactual outcome of compliers:

$$\begin{aligned} W(\pi) &= E[Y(D(0))] + E[(Y(1) - Y(0))(D(1) - D(0)) \cdot \pi(X)] \\ &= E[Y(D(0))] + E[\Delta_c(X)\varrho_c(X) \cdot \pi(X)]. \end{aligned}$$

For always-takers and never-takers, no encouragement rule can alter their outcomes.

We proceed to present two extensions analogous to those in Section 2.4. In B.3.1, we allow for multiple instruments. In B.3.2, we incorporate budget constraints.

B.3.1 Multiple Instruments

Consider a setting in which there are L instruments available. Let Z_1 be the binary instrument that can be manipulated, $Z_{-1} \equiv (Z_2, \ldots, Z_L)^\top \in \mathcal{Z}_{-1} \subset \mathbb{R}^{L-1}$ be a vector of additional instruments, and $Z = (Z_1, Z_{-1}^\top)^\top$. An encouragement rule is a mapping $\pi : \mathcal{X} \times \mathcal{Z}_{-1} \to \{0, 1\}$ that determines the manipulated value of Z_1 while leaving Z_{-1} unchanged. Define the social welfare criterion as $W(\pi) = E[Y(D_1(\pi(X, Z_{-1})))]$. Under (2.4) and Assumption 2.5, $W(\pi)$ is identified as

$$W(\pi) = E[E[Y|X, Z_{-1}, Z_1 = 1] \cdot \pi(X, Z_{-1}) + E[Y|X, Z_{-1}, Z_1 = 0] \cdot (1 - \pi(X, Z_{-1}))].$$

The additional instruments Z_{-1} are treated equivalently to covariates X.

In the case of two binary instruments, to understand which subpopulation benefits from the encouragement rule, we impose further assumptions on the selection behavior:

Assumption B.2 (Component-Wise Increasing Propensity Score)**.** *For each $x \in \mathcal{X}$ and $z_1, z_2 \in \{0, 1\}$, $p(x, 1, z_2) \geq p(x, 0, z_2)$ and $p(x, z_1, 1) \geq p(x, z_1, 0)$.*

Under (2.4) and Assumption B.2, we can partition the population into the six compliance groups presented in Table B.1. Denote by $G \in \mathcal{G} = \{\text{nt}, \text{at}, \text{1c}, \text{2c}, \text{ec}, \text{rc}\}$ the compliance group identity. For each $g \in \mathcal{G}$ and $z_2 \in \{0, 1\}$, let $\varrho_g(x, z_2) = \Pr(G =$

$g|X = x, Z_2 = z_2)$ be the conditional population share of compliance group g and $\Delta_g(x, z_2) = E[Y(1) - Y(0)|X = x, Z_2 = z_2, G = g]$ be the compliance-group-specific CATE. Then, the social welfare criterion can also be written as

$$
\begin{aligned}
W(\pi) &= E[Y(D_1(0))] + E[(Y(1) - Y(0))(D_1(1) - D_1(0)) \cdot \pi(X, Z_2)] \\
&= E[Y(D_1(0))] \\
&\quad + E[E[(Y(1) - Y(0))(D(1,1) - D(0,1))|X, Z_2 = 1] \cdot \Pr(Z_2 = 1|X) \cdot \pi(X, 1)] \\
&\quad + E[E[(Y(1) - Y(0))(D(1,0) - D(0,0))|X, Z_2 = 0] \cdot \Pr(Z_2 = 0|X) \cdot \pi(X, 0)] \\
&= E[Y(D_1(0))] \\
&\quad + E[(\varrho_{1c}(X,1)\Delta_{1c}(X,1) + \varrho_{rc}(X,1)\Delta_{rc}(X,1)) \cdot \Pr(Z_2 = 1|X) \cdot \pi(X, 1)] \\
&\quad + E[(\varrho_{1c}(X,0)\Delta_{1c}(X,0) + \varrho_{ec}(X,0)\Delta_{ec}(X,0)) \cdot \Pr(Z_2 = 0|X) \cdot \pi(X, 0)].
\end{aligned}
$$

Therefore, besides Z_1 compliers, the encouragement rule would affect the outcomes of reluctant compliers with $Z_2 = 1$ and eager compliers with $Z_2 = 0$.

Table B.1: Compliance Groups (Mogstad et al., 2021, Proposition 4)

Name	$D(0,0)$	$D(0,1)$	$D(1,0)$	$D(1,1)$
Never-takers (nt)	N	N	N	N
Always-takers (at)	T	T	T	T
Z_1 compliers (1c)	N	N	T	T
Z_2 compliers (2c)	N	T	N	T
Eager compliers (ec)	N	T	T	T
Reluctant compliers (rc)	N	N	N	T

Notes: "T" indicates treatment, and "N" indicates non-treatment.

B.3.2 Budget Constraints

Suppose that the policymaker faces a harsh budget constraint such that the proportion of the population receiving treatment cannot exceed $\kappa \in (0, 1)$. Then we face a

constrained optimization problem:

$$\max_{\pi \in \Pi} W(\pi) \text{ s.t. } B(\pi) \leq \kappa,$$

where $B(\pi) = E[D(\pi(X))]$. Let π_B^* denote the solution. We maintain (2.2) and Assumptions 2.1 and B.1 for the rest of this subsection.[1]

When Π is unrestricted, π_B^* coincides with the optimal deterministic individualized encouragement rule (IER) studied by Qiu et al. (2021). Provided that the distribution of $\Delta_c(X)$ is continuous, π_B^* admits a closed form: $\pi_B^*(x) = 1\{\Delta_c(x) \geq \max\{\underline{\Delta}, 0\}\}$, where $\underline{\Delta} = \inf\{\Delta \in \mathbb{R} : E[1\{\Delta_c(X) \geq \Delta\}\varrho_c(X)] \leq \kappa - E[D(0)]\}$. Intuitively, π_B^* assigns encouragement to compliers in decreasing order of $\Delta_c(x)$ until the resources are exhausted.

One simple idea is to enforce random rationing as in Kitagawa and Tetenov (2018b): if π violates the resource constraint, then the encouragement is randomly allocated to a fraction $\frac{\kappa - E[D(0)]}{E[D(\pi(X))] - E[D(0)]}$ of individuals with $\pi(X) = 1$ independently of everything else. Define the resource-constrained welfare criterion as

$$W^\kappa(\pi) = E[Y(D(0))] + E[(Y(D(1)) - Y(D(0))) \cdot \pi(X)] \cdot \min\left\{1, \frac{\kappa - E[D(0)]}{E[D(\pi(X))] - E[D(0)]}\right\}.$$

The optimal encouragement rule under random rationing solves $\max_{\pi \in \Pi} W^\kappa(\pi)$, which uses the resources less efficiently than π_B^*.

B.4 Sup-norm Convergence Rate in Expectation for Nonparametric Propensity Score Estimators

We consider the nonparametric regression model

$$D = p(\tilde{X}) + \epsilon, \quad E[\epsilon | \tilde{X}] = 0,$$

[1] We implicitly assume that $E[D(0)] \leq \kappa$.

where $\tilde{X} = (X^\top, Z)^\top \in \tilde{\mathcal{X}} \subset \mathbb{R}^{d_{\tilde{x}}}$. To force the resulting estimator $\hat{p}(\tilde{x})$ to lie between 0 and 1, one can use a trimmed version as in Carneiro and Lee (2009, Eq. (4.2)).

B.4.1 Local Polynomial Estimators

We follow Audibert and Tsybakov (2007) to impose the following restrictions:

Assumption B.3 (Local Polynomial Estimators). *(i) $p(\cdot)$ belongs to a Hölder class of functions with degree $s \geq 1$ and constant $0 < R < \infty$. (ii) Let $\mathrm{Leb}(\cdot)$ be the Lebesgue measure on $\mathbb{R}^{d_{\tilde{x}}}$. There exist constants c_0 and r_0 such that*

$$\mathrm{Leb}(\tilde{\mathcal{X}} \cap B(\tilde{x}, r)) \geq c_0 \,\mathrm{Leb}(B(\tilde{x}, r)) \quad \forall 0 < r \leq r_0,\ \forall \tilde{x} \in \tilde{\mathcal{X}},$$

where $B(\tilde{x}, r)$ denotes the closed Euclidean ball in $\mathbb{R}^{d_{\tilde{x}}}$ centered at $\tilde{x}$ and of radius r. Moreover, the marginal distribution of $\tilde{X}$ has the density function f with respect to the Lebesgue measure of $\mathbb{R}^{d_{\tilde{x}}}$ such that $0 < f_{\min} \leq f(\tilde{x}) \leq f_{\max} < \infty$ $\forall \tilde{x} \in \tilde{\mathcal{X}}$. (iii) The kernel function $K(\cdot)$ satisfies: $\exists c > 0$ such that $K(u) \geq c1\{\|u\| \leq c\}$ $\forall u \in \mathbb{R}^{d_{\tilde{x}}}$, $\int_{\mathbb{R}^{d_{\tilde{x}}}} K(u)du = 1$, $\int_{\mathbb{R}^{d_{\tilde{x}}}} (1 + \|u\|^{4s}) K^2(u)du < \infty$, and $\sup_{u \in \mathbb{R}^{d_{\tilde{x}}}} (1 + \|u\|^{2s}) K(u) < \infty$.

We consider the local polynomial regression fit for $p(\tilde{x})$ with degree $l = s - 1$:

$$\begin{aligned} \hat{p}(\tilde{x}) &= (\hat{\xi}(\tilde{x}))^\top U(0) \cdot 1\{\lambda_{\min}(B(\tilde{x})) \geq (\log n)^{-1}\}, \\ \hat{\xi}(\tilde{x}) &= \arg\min_{\xi} \sum_{i=1}^{n} \Big[D_i - \xi^\top U\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big)\Big]^2 K\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big), \end{aligned} \tag{B.11}$$

where $U\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big)$ is a vector with elements indexed by the multi-index $(\ell_1, \ldots, \ell_{d_{\tilde{x}}}) \in \mathbb{N}^{d_{\tilde{x}}}$:

$$U\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big) = \Big\{ \Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big)_1^{\ell_1} \cdots \Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big)_{d_{\tilde{x}}}^{\ell_{d_{\tilde{x}}}} \Big\}_{0 \leq \sum_{i=1}^{d_{\tilde{x}}} \ell_i \leq l},$$

and $\lambda_{\min}(B(\tilde{x}))$ is the smallest eigenvalue of

$$B(\tilde{x}) = \frac{1}{nh^{d_{\tilde{x}}}} \sum_{i=1}^{n} U\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big) \Big(U\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big)\Big)^\top K\Big(\frac{\tilde{X}_i - \tilde{x}}{h}\Big).$$

Theorem B.1 states the sup-norm convergence rate in expectation for $\hat{p}$. It is straightforward to calculate the fastest convergence rate in Assumption 2.4(ii) as $\psi_n =$

$n^{\frac{1}{2}-\frac{1}{2+4s/d_{\tilde{x}}}}$.

Theorem B.1. *Suppose that Assumption B.3 holds. Then, for the local polynomial fit $\hat{p}(\cdot)$ for $p(\cdot)$ defined in (B.11), we have*

$$E\Big[\sup_{\tilde{x}\in\tilde{\mathcal{X}}}|\hat{p}(\tilde{x})-p(\tilde{x})|^2\Big]^{1/2}=O\Big(h^s+\frac{1}{\sqrt{nh^{d_{\tilde{x}}}}}\Big).$$

Proof. By Theorem 3.2 of Audibert and Tsybakov (2007), there exist positive constants C_1, C_2, C_3 that only depend only on s, R, $d_{\tilde{x}}$, c_0, r_0, $f_{\min}$, $f_{\max}$, and the kernel K, such that, for any $0<h<r_0/c_0$, and $C_3h^s<\delta$, and any $n\geq 1$,

$$\Pr\Big(\sup_{\tilde{x}\in\tilde{\mathcal{X}}}|\hat{p}(\tilde{x})-p(\tilde{x})|\geq\delta\Big)\leq C_1\exp(-C_2nh^{d_{\tilde{x}}}\delta^2).$$

It follows that

$$\begin{aligned}E\Big[\sup_{\tilde{x}\in\tilde{\mathcal{X}}}|\hat{p}(\tilde{x})-p(\tilde{x})|^2\Big]&=\int_0^\infty\Pr\Big(\sup_{\tilde{x}\in\tilde{\mathcal{X}}}|\hat{p}(\tilde{x})-p(\tilde{x})|^2\geq\delta\Big)d\delta\\&\leq C_3^2h^{2s}+C_1\int_0^\infty\exp(-C_2nh^{d_{\tilde{x}}}\delta)d\delta\\&=C_3^2h^{2s}+\frac{C_4}{nh^{d_{\tilde{x}}}},\end{aligned}$$

where $C_4=C_1\int_0^\infty\exp(-C_2\delta')d\delta'<\infty$. □

B.4.2 Series Estimators

Let us introduce some notations. Let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of a matrix. Let the exponent $^-$ denote the Moore–Penrose generalized inverse. Let $\|\cdot\|$ denote the Euclidean norm when applied to vectors and the matrix spectral norm (i.e., the largest singular value) when applied to matrices. Let $\|\cdot\|_\infty$ denote the sup norm, i.e., if $f:\tilde{\mathcal{X}}\to\mathbb{R}$ then $\|f\|_\infty=\sup_{\tilde{x}\in\tilde{\mathcal{X}}}|f(\tilde{x})|$.

We consider the series least-squares estimator of p:

$$\hat{p}(\tilde{x})=b^k(\tilde{x})^\top(B^\top B)^-B^\top\mathbf{D},$$

where $b_{k1}, \ldots, b_{kk}$ are a collection of k sieve basis functions, and

$$b^k(\tilde{x}) = (b_{k1}(\tilde{x}), \ldots, b_{kk}(\tilde{x}))^\top, \quad B = (b^k(\tilde{X}_1), \ldots, b^k(\tilde{X}_n))^\top, \quad \mathbf{D} = (D_1, \ldots, D_n)^\top.$$

Define $\zeta_k = \sup_{\tilde{x} \in \tilde{\mathcal{X}}} \|b^k(\tilde{x})\|$ and $\lambda_k = [\lambda_{\min}(E[b^k(\tilde{X})b^k(\tilde{X})^\top])]^{-1/2}$. Let $B_k = \text{clsp}\{b_{k1}, \ldots, b_{kk}\}$ denote the closed linear span of the basis functions. We impose the following regularity conditions:

Assumption B.4 (Series Estimators)**.** *(i)* $\lambda_{\min}(E[b^k(\tilde{X})b^k(\tilde{X})^\top]) > 0$ *for each* k*. (ii) There exist* $\nu, \tilde{c} > 0$ *such that* $\sup_{\tilde{x} \in \tilde{\mathcal{X}}} E[|\epsilon|^q | \tilde{X} = \tilde{x}] \leq \frac{q!}{2}\nu^2 \tilde{c}^{q-2}$ *for all* $q \geq 2$*. (iii) There exists* $\rho > 0$ *such that* $\inf_{h \in B_k} \|p - h\|_\infty = O(k^{-\rho})$ *as* $k \to \infty$*.*

Let $\tilde{b}^k(\tilde{x})$ denote the orthonormalized vector of basis functions, namely

$$\tilde{b}^k(\tilde{x}) = E[b^k(\tilde{X})b^k(\tilde{X})^\top]^{-1/2} b^k(\tilde{x})$$

and let $\tilde{B} = (\tilde{b}^k(\tilde{X}_1), \ldots, \tilde{b}^k(\tilde{X}_n))^\top$. Let $\tilde{p}$ denote the projection of p onto B_k under the empirical measure, that is,

$$\tilde{p}(\tilde{x}) = b^k(\tilde{x})^\top (B^\top B)^- B^\top \mathbf{P} = \tilde{b}^k(\tilde{x})^\top (\tilde{B}^\top \tilde{B})^- \tilde{B}^\top \mathbf{P},$$

where $\mathbf{P} = (p(\tilde{X}_1), \ldots, p(\tilde{X}_n))^\top$. We can bound $\|\hat{p} - p\|_\infty$ using

$$\begin{aligned} \|\hat{p} - p\|_\infty &\leq \|p - \tilde{p}\|_\infty + \|\hat{p} - \tilde{p}\|_\infty \\ &= \text{bias term} + \text{variance term}. \end{aligned}$$

We state three preparatory lemmas. Lemma B.1 gives an exponential tail bound for $\|\tilde{B}^\top \tilde{B}/n - I_k\|$. Lemma B.2 provides an exponential tail bound on the sup-norm variance term. Lemma B.3 provides a bound on the sup-norm bias term. The proofs are in Appendix B.7.

Lemma B.1. *Under Assumption B.4(i), we have*

$$\Pr(\|\tilde{B}^\top \tilde{B}/n - I_k\| \geq \frac{1}{2}) \leq 2k \exp\Big(- \frac{C_5 n}{\zeta_k^2 \lambda_k^2 + 1}\Big)$$

for some finite positive constant C_5.

Let $\mathcal{A}_n$ denote the event on which $\|\tilde{B}^\top \tilde{B}/n - I_k\| \leq \frac{1}{2}$. Let $1_{\mathcal{A}_n}$ denote the indicator function of $\mathcal{A}_n$.

Lemma B.2. *Under Assumptions B.4(i) and (ii), for any $\delta \in (0, 1]$, we have*

$$\Pr(1_{\mathcal{A}_n} \|\hat{p} - \tilde{p}\|_\infty \geq \delta) \leq \exp\Big(- \frac{C_6 n \delta^2}{\zeta_k^2 \lambda_k^2}\Big)$$

for some finite positive constant C_6.

Let $P_{k,n}$ be the empirical projection operator onto B_k, namely

$$P_{k,n} h(\tilde{x}) = b^k(\tilde{x})^\top (B^\top B)^- B^\top \mathbf{H} = \tilde{b}^k(\tilde{x})^\top (\tilde{B}^\top \tilde{B})^- \tilde{B}^\top \mathbf{H}$$

for any $h : \tilde{\mathcal{X}} \to \mathbb{R}$, where $\mathbf{H} = (h(\tilde{X}_1), \ldots, h(\tilde{X}_n))^\top$. Let $L^\infty(\tilde{X})$ denote the space of bounded functions under the sup norm. Let

$$\|P_{k,n}\|_\infty = \sup_{h \in L^\infty(\tilde{X}) : \|h\|_\infty \neq 0} \frac{\|P_{k,n} h\|_\infty}{\|h\|_\infty}.$$

Lemma B.3. *Under Assumption B.4(i), we have*

$$\|\tilde{p} - p\|_\infty \leq (1 + \|P_{k,n}\|_\infty) \inf_{h \in B_k} \|p - h\|_\infty$$

and $1_{\mathcal{A}_n} \|P_{k,n}\|_\infty \leq \sqrt{2} \zeta_k \lambda_k$.

Theorem B.2 concludes the sup-norm convergence rate in expectation for $\hat{p}$.

Theorem B.2. *Under Assumption B.4, we have*

$$E[\|\hat{p} - p\|_\infty^2]^{1/2} = O(\zeta_k \lambda_k (k^{-\rho} + \sqrt{k}/\sqrt{n})).$$

Proof. Combining Lemmas B.1–B.3, there exist positive constants C_7, C_8, and C_9 such that if $C_7\zeta_k\lambda_k k^{-\rho} \le \delta \le 1$,

$$\begin{aligned}\Pr(\|\hat{p}-p\|_\infty \ge \delta) &\le 1 - \Pr(\mathcal{A}_n) + \Pr(1_{\mathcal{A}_n}\|\hat{p}-p\|_\infty \ge \delta) \\ &\le \Pr(\|\tilde{B}^\top\tilde{B}/n - I_k\| \ge \frac{1}{2}) + \Pr(1_{\mathcal{A}_n}\|\hat{p}-\tilde{p}\|_\infty \ge \frac{\delta}{2}) \\ &\le C_8 k \exp\Big(-\frac{C_9 n\delta^2}{\zeta_k^2\lambda_k^2}\Big)\end{aligned}$$

for all sufficiently large k. For $\delta > 1$, the same inequality holds since $\hat{p}$ and p take values in $[0,1]$. It follows that there exist positive constants C_{10} and C_{11} such that

$$\begin{aligned}E[\|\hat{p}-p\|_\infty^2] &= \int_0^\infty \Pr(\|\hat{p}-p\|_\infty^2 \ge \delta)d\delta \\ &\le C_{10}\zeta_k^2\lambda_k^2 k^{-2\rho} + C_8 k\int_0^\infty \exp\Big(-\frac{C_9 n\delta}{\zeta_k^2\lambda_k^2}\Big)d\delta \\ &= C_{10}\zeta_k^2\lambda_k^2 k^{-2\rho} + C_{11}\zeta_k^2\lambda_k^2\frac{k}{n}\end{aligned}$$

for all sufficiently large k. □

When $\tilde{\mathcal{X}}$ is compact and rectangular, and X has a probability density function that is bounded away from zero on $\tilde{\mathcal{X}}$, $\zeta_k\lambda_k = O(\sqrt{k})$ for regression splines and $\zeta_k\lambda_k = O(k)$ for polynomials. If we further assume that $p(\cdot)$ is continuously differentiable of order s on $\tilde{\mathcal{X}}$, then Assumption B.4(iii) holds with $\rho = s/d_{\tilde{x}}$. It is straightforward to calculate the fastest convergence rate in Assumption 2.4(ii) as $\psi_n = n^{\frac{1}{2}-\frac{1}{1+2s/d_{\tilde{x}}}}$ for regression splines and $\psi_n = n^{\frac{1}{2}-\frac{3}{2+4s/d_{\tilde{x}}}}$ for polynomials.

B.5 Sup-norm Convergence Rate in Expectation for Parametric MTE Estimators

We consider a parametric regression function for the MTE:

$$E[Y|p(X,Z)=u,X=x]=x^\top\beta_0+x^\top(\beta_1-\beta_0)u+\sum_{j=2}^{J}\eta_j u^j.$$

Let $W=((1-p(X,Z))X^\top,p(X,Z)X^\top,p(X,Z)^2,\ldots,p(X,Z)^J)^\top$ and $\vartheta=(\beta_0^\top,\beta_1^\top,\eta_2,\ldots,\eta_J)^\top$. Then, ϑ is identified as $\vartheta=E[WW^\top]^{-1}E[WY]$ provided that $E[WW^\top]$ is positive definite. Given an estimator $\hat{p}(x,z)$ for $p(x,z)$, define the regressor as

$$\hat{W}=((1-\hat{p}(X,Z))X^\top,\hat{p}(X,Z)X^\top,\hat{p}(X,Z)^2,\ldots,\hat{p}(X,Z)^J)^\top.$$

The OLS estimator for ϑ is obtained by regressing Y on $\hat{W}$:

$$\hat{\vartheta}=(\hat{\beta}_0^\top,\hat{\beta}_1^\top,\hat{\eta}_2,\ldots,\hat{\eta}_J)^\top=E_n[\hat{W}\hat{W}^\top]^{-1}E_n[\hat{W}Y].$$

Then, the MTE can be estimated by

$$\widehat{\mathrm{MTE}}(u,x)=x^\top(\hat{\beta}_1-\hat{\beta}_0)+\sum_{j=2}^{J}j\hat{\eta}_j u^{j-1}.$$

For this concrete estimator, we provide primitive conditions that guarantee the high-level condition in Assumption 2.4(iii) to hold.

Assumption B.5 (Polynomial MTE Model)**.** *Let C and c be positive constants. (i) X and Y have bounded support. (ii) The parameter space for ϑ is compact so that for sufficiently large n, $\|\hat{\vartheta}\|+\|\vartheta\|\leq C$ almost surely. (iii) $\lambda_{\min}(E[WW^\top])\geq c$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix.*

Proposition B.1. *Under Assumptions 2.4(ii) and B.5, Assumption 2.4(iii) holds with $\phi_n=\psi_n$.*

Proof. We have

$$
\begin{aligned}
&E\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|\widehat{\mathrm{MTE}}(u,X)-\mathrm{MTE}(u,X)|\Big]\Big] \\
&= E\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}\Big|X^\top(\hat{\beta}_1-\hat{\beta}_0)+\sum_{j=2}^{J}j\hat{\eta}_j u^{j-1}-X^\top(\beta_1-\beta_0)-\sum_{j=2}^{J}j\eta_j u^{j-1}\Big|\Big]\Big] \\
&\le \sum_{d\in\{0,1\}}E[E_n[|X^\top(\hat{\beta}_d-\beta_d)|]]+\sum_{j=2}^{J}jE\Big[E_n\Big[\sup_{u\in\bar{\mathcal{U}}(X)}|(\hat{\eta}_j-\eta_j)u^{j-1}|\Big]\Big] \\
&\le \sum_{d\in\{0,1\}}E[E_n[\|X\|]\|\hat{\beta}_d-\beta_d\|]+\sum_{j=2}^{J}jE[|\hat{\eta}_j-\eta_j|],
\end{aligned}
$$

where the first inequality follows from the triangle inequality and the second inequality follows from the Cauchy-Schwarz inequality. Given Assumption B.5(i), it suffices to show

$$E[\|\hat{\vartheta}-\vartheta\|]=O(\psi_n^{-1}).$$

We can write

$$
\begin{aligned}
E[WW^\top](\hat{\vartheta}-\vartheta) = {}& E_n[(\hat{W}-W)Y]-E_n[\hat{W}\hat{W}^\top-WW^\top]\hat{\vartheta} \\
&+(E_n-E)[WY]-(E_n-E)[WW^\top]\hat{\vartheta}.
\end{aligned}
$$

Let $\|\cdot\|$ denote the matrix spectral norm when applied to matrices. There exist positive constants $C_1, C_2<\infty$ such that

$$
\begin{aligned}
\|(\hat{W}-W)Y\| &\le \|(\hat{W}-W)\||Y| \le |Y|(C_1+C_2\|X\|)|\hat{p}(X,Z)-p(X,Z)|, \\
\|\hat{W}\hat{W}^\top-WW^\top\| &\le \|(\hat{W}-W)\|\|(\hat{W}+W)\| \le (C_1+C_2\|X\|^2)|\hat{p}(X,Z)-p(X,Z)|.
\end{aligned}
$$

Then, by Assumptions 2.4(ii) and B.5(i),

$$E[E_n[\|(\hat{W}-W)Y\|]]=O(\psi_n^{-1}), \tag{B.12}$$

$$E[E_n[\|\hat{W}\hat{W}^\top-WW^\top\|]]=O(\psi_n^{-1}). \tag{B.13}$$

Also, by Assumption B.5(i),

$$E[\|(E_n - E)[WY]\|] = O(n^{-1/2}), \tag{B.14}$$
$$E[\|(E_n - E)[WW^\top]\|] = O(n^{-1/2}). \tag{B.15}$$

By the triangle inequality,

$$\begin{aligned} E[\|\hat{\vartheta} - \vartheta\|] \leq\ & \lambda_{\min}(E[WW^\top])^{-1} E[E_n[\|(\hat{W} - W)Y\|]] \\ & + \lambda_{\min}(E[WW^\top])^{-1} E[E_n[\|\hat{W}\hat{W}^\top - WW^\top\|]\|\hat{\vartheta}\|] \\ & + \lambda_{\min}(E[WW^\top])^{-1} E[\|(E_n - E)[WY]\|] \\ & + \lambda_{\min}(E[WW^\top])^{-1} E[\|(E_n - E)[WW^\top]\|\|\hat{\vartheta}\|], \end{aligned}$$

where the right-hand side is $O(\psi_n^{-1})$ by Assumptions B.5(ii)–(iii) and (B.12)–(B.15). □

B.6 Doubly Robust Approach

The idea of the doubly robust approach of Athey and Wager (2021) is to use an alternative objective for policy learning that consists of doubly robust scores. A natural candidate for the doubly robust score arises from studying the influence function of the population welfare criterion. We focus on the nonparametric identification of $W(\pi)$ and impose the following assumptions:

Assumption B.6 (Nonparametric Identification of $W(\pi)$)**.** *(i) For each $d \in \{0, 1\}$, $\mathrm{Supp}(X, \alpha_d(Z)) \subset \mathrm{Supp}(X, Z)$. (ii) The conditional distribution of $p(X, Z)$ given X is absolutely continuous with respect to the Lebesgue measure.*

As discussed in Example 2.1, under Assumptions 2.1 and B.6, we have

$$W(\pi) = E[Y(D(\alpha_0(Z)))] + E[\pi(X, Z) \cdot (\varphi(X, p(X, \alpha_1(Z)); p) - \varphi(X, p(Z, \alpha_0(Z)); p))],$$

where $\varphi(x, u; p) = E[Y|X = x, p(X, Z) = u]$. Fix $\pi \in \Pi$ and define

$$m(x, z; p, \varphi) = \pi(x, z)(\varphi(x, p(x, \alpha_1(z)); p) - \varphi(x, p(x, \alpha_0(z)); p))$$

so that $W(\pi) = E[Y(D(\alpha_0(Z)))] + E[m(X, Z; p, \varphi)]$. Lemma B.4 calculates the influence function of $E[m(X, Z; p, \varphi)]$ following Ichimura and Newey (2022). A proof is provided in Appendix B.7. It is worth noting that the estimation error of p has two contributions to the influence function of $E[m(X, Z; p, \varphi)]$ through φ. First, the estimate of p serves as a generated regressor that changes the conditional expectation estimator. Second, the estimate of p enters as the argument at which the conditional expectation estimator is evaluated. We adopt the approach from Hahn and Ridder (2013) to show that the two contributions cancel each other.

Lemma B.4. *Let $F_\tau = (1-\tau)F_0 + \tau H, 0 < \tau < 1$ denote a convex combination of the true CDF F_0 with another CDF H. Let $p_\tau = p(F_\tau)$ and $\varphi_\tau = \varphi(F_\tau)$. Then,*

$$\frac{\partial}{\partial \tau} E[m(X, Z; p_\tau, \varphi_\tau)] = E_H[\pi(X, Z) g(X, Z)(Y - \varphi_0(X, p_0(X, Z); p_0))],$$

where

$$g(x, z) = \frac{f_{X,\alpha_1(Z)}(x, z) - f_{X,\alpha_0(Z)}(x, z)}{f_{X,Z}(x, z)}.$$

By Lemma B.4, we can construct the doubly robust score in the following steps.

1. Divide the data into K evenly-sized folds. For each fold $k = 1, \ldots, K$, use the other $K - 1$ data folds to

   (a) estimate $p(x, z)$ and $g(x, z)$; denote the resulting estimates by $\hat{p}^{(-k)}(x, z)$ and $\hat{g}^{(-k)}(x, z)$;

   (b) estimate $\varphi(x, u; \hat{p}^{(-k)})$; denote the resulting estimate by $\varphi^{(-k)}(x, u; \hat{p}^{(-k)})$.

2. Calculate the doubly robust score as

$$\begin{aligned}\hat{\Gamma}_i = {} & \hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, \alpha_1(Z_i)); \hat{p}^{(-k(i))}) \\ & - \hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, \alpha_0(Z_i)); \hat{p}^{(-k(i))}) \\ & + \hat{g}^{(-k(i))}(X_i, Z_i)(Y_i - \hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, Z_i); \hat{p}^{(-k(i))})),\end{aligned}$$

where $k(i) \in \{1, \ldots, K\}$ denotes the fold containing the ith observation.

Define the *doubly-robust encouragement rule* as

$$\hat{\pi}_{\mathrm{DR}} \in \arg\max_{\pi \in \Pi} \hat{W}_n^{\mathrm{DR}}(\pi), \quad \hat{W}_n^{\mathrm{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \pi(X_i, Z_i) \hat{\Gamma}_i. \tag{B.16}$$

To analyze the regret of $\hat{\pi}_{\mathrm{DR}}$, we impose the following assumptions:

Assumption B.7 (Doubly-Robust Encouragement Rule)**.** *(i)* $\sup_{x,z} |g(x,z)| < \infty$, $\sup_{x,z} |\varphi(x, p(x,z); p)| < \infty$, *and for* $d \in \{0,1\}$, $\sup_{x,z} |\varphi(x, p(x, \alpha_d(z)); p)| < \infty$. *(ii)* $\varepsilon = Y - \varphi(X, p(X,Z); p)$ *is uniformly sub-Gaussian conditionally on* (X,Z) *and has second moments uniformly bounded from below. (iii)* $\sup_{x,z} |\hat{g}^{(-k)}(x,z)| < \infty$, $\sup_{x,z} |\hat{\varphi}^{(-k)}(x, \hat{p}^{(-k)}(x,z); \hat{p}^{(-k)})| < \infty$, *and for* $d \in \{0,1\}$, $\sup_{x,z} |\hat{\varphi}^{(-k)}(x, \hat{p}^{(-k)}(x, \alpha_d(z)); \hat{p}^{(-k)})| < \infty$ *almost surely. (iv) There exist* $0 < \zeta_g, \zeta_\varphi < 1$ *with* $\zeta_g + \zeta_\varphi \geq 1$ *and* $a(n) \to 0$ *such that*

$$\begin{aligned} E[(\hat{g}^{(-k(i))}(X_i, Z_i) - g(X_i, Z_i))^2] &\leq \frac{a(n)}{n^{\zeta_g}}, \\ E[(\hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, Z_i); \hat{p}^{(-k(i))}) - \varphi(X_i, p(X_i, Z_i); p))^2] &\leq \frac{a(n)}{n^{\zeta_\varphi}}, \end{aligned}$$

and for $d \in \{0,1\}$,

$$E[(\hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, \alpha_d(Z_i)); \hat{p}^{(-k(i))}) - \varphi(X_i, p(X_i, \alpha_d(Z_i)); p))^2] \leq \frac{a(n)}{n^{\zeta_\varphi}}.$$

Theorem B.3 shows that the average regret of $\hat{\pi}_{\mathrm{DR}}$ decays no slower than $n^{-1/2}$.

Theorem B.3. *Under Assumptions 2.1, 2.3(ii), B.6, and B.7, we have*

$$E[R(\hat{\pi}_{\mathrm{DR}})] = O(n^{-1/2}).$$

Proof. We follow Athey and Wager (2021) to work with

$$\begin{aligned} A(\pi) &= E[(2\pi(X,Z) - 1) \cdot (\varphi(X, p(X, \alpha_1(Z)); p) - \varphi(X, p(Z, \alpha_0(Z)); p))], \\ \hat{A}_n(\pi) &= \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i, Z_i) - 1) \hat{\Gamma}_i. \end{aligned}$$

Note that $\max_{\pi' \in \Pi} A(\pi') - A(\pi) = 2R(\pi)$. It is convenient to define an ideal version of the objective in (B.16) based on the true influence scores:

$$\begin{aligned}\tilde{A}_n(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i, Z_i) - 1)\Gamma_i, \\ \Gamma_i &= \varphi(X_i, p(X_i, \alpha_1(Z_i)); p) - \varphi(X_i, p(X_i, \alpha_0(Z_i)); p) \\ &\quad + g(X_i, Z_i)(Y_i - \varphi(X_i, p(X_i, Z_i); p)).\end{aligned}$$

By writing

$$\hat{A}_n(\pi) - A(\pi) = \hat{A}_n(\pi) - \tilde{A}_n(\pi) + \tilde{A}_n(\pi) - A(\pi),$$

we study stochastic fluctuations of $\hat{A}_n(\pi) - A(\pi)$ for $\pi \in \Pi$ in two steps. First, we bound $|\tilde{A}_n(\pi) - A(\pi)|$ over λ-slices of Π defined as

$$\Pi^\lambda = \{\pi \in \Pi : R(\pi) \leq \lambda\}.$$

By Assumptions B.7(i)–(ii) and Hoeffding's inequality, the Γ_i are uniformly sub-Gaussian and have variance bounded from below. Hence, Corollary 3 of Athey and Wager (2021) holds with $S_n = E[\Gamma^2]$ and $S_n^\lambda = \sup\{\mathrm{Var}[(2\pi(X, Z) - 1)\Gamma] : \pi \in \Pi^\lambda\}$. Next, we bound $|\hat{A}_n(\pi) - \tilde{A}_n(\pi)|$ over $\pi \in \Pi$. To save space, we write

$$\begin{aligned}\Delta\varphi_i &= \varphi(X_i, p(X_i, \alpha_1(Z_i)); p) - \varphi(X_i, p(X_i, \alpha_0(Z_i)); p), \\ \Delta\hat{\varphi}_i &= \hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, \alpha_1(Z_i)); \hat{p}^{(-k(i))}) - \hat{\varphi}^{(-k(i))}(X_i, \hat{p}^{(-k(i))}(X_i, \alpha_0(Z_i)); \hat{p}^{(-k(i))}).\end{aligned}$$

For any fixed π, we can expand $\hat{A}_n(\pi) - \tilde{A}_n(\pi)$ as

$$\hat{A}_n(\pi) - \tilde{A}_n(\pi) = D_1(\pi) + D_2(\pi) - D_3(\pi),$$

where

$$
\begin{aligned}
D_1(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i,Z_i)-1)(Y_i-\varphi(X_i,p(X_i,Z_i);p))(\hat{g}^{(-k(i))}(X_i,Z_i)-g(X_i,Z_i)),\\
D_2(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i,Z_i)-1)[\Delta\hat{\varphi}_i-\Delta\varphi_i\\
&\quad -g(X_i,Z_i)(\hat{\varphi}^{(-k(i))}(X_i,\hat{p}^{(-k(i))}(X_i,Z_i);\hat{p}^{(-k(i))})-\varphi(X_i,p(X_i,Z_i);p))],\\
D_3(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i,Z_i)-1)(\hat{\varphi}^{(-k(i))}(X_i,\hat{p}^{(-k(i))}(X_i,Z_i);\hat{p}^{(-k(i))})-\varphi(X_i,p(X_i,Z_i);p))\\
&\quad \cdot(\hat{g}^{(-k(i))}(X_i,Z_i)-g(X_i,Z_i)).
\end{aligned}
$$

We bound all 3 summands separately. We start with $D_1(\pi)$. It is helpful to separate out the contributions of the K different folds:

$$
D_1^{(k)}(\pi) = \frac{1}{n_k}\sum_{i:k(i)=k}(2\pi(X_i,Z_i)-1)(Y_i-\varphi(X_i,p(X_i,Z_i);p))(\hat{g}^{(-k(i))}(X_i,Z_i)-g(X_i,Z_i))
$$

so that $D_1(\pi)=\sum_{k=1}^{K}\frac{n_k}{n}D_1^{(k)}(\pi)$, where $n_k=|\{i:k(i)=k\}|$ denotes the number of observations in the kth fold. Note that $E[Y_i-\varphi(X_i,p(X_i,Z_i);p)|X_i,Z_i,\hat{g}^{(-k(i))}(\cdot)]=0$. Hence, conditional on $\hat{g}^{(-k)}(\cdot)$ fit on the other $K-1$ folds, $D_1^{(k)}(\pi)$ has zero mean and asymptotic variance

$$
V_1(k) = E[(\hat{g}^{(-k(i))}(X_i,Z_i)-g(X_i,Z_i))^2\,\mathrm{Var}[Y_i-\varphi(X_i,p(X_i,Z_i);p)|X_i,Z_i]|\hat{g}^{(-k)}(\cdot)].
$$

By Assumptions B.7(i)–(iii), we can apply Corollary 3 of Athey and Wager (2021) to establish that

$$
E\Big[\sup_{\pi\in\Pi}\big|D_1^{(k)}(\pi)\big|\,|\hat{g}^{(-k)}(\cdot)\Big] = O\Big(\sqrt{\mathrm{VC}(\Pi)\frac{V_1(k)}{n_k}}\Big). \tag{B.17}
$$

Since for a finite number of evenly-sized folds, $n_k/n\to K^{-1}$, we can use Assumption B.7(iv) to check that

$$
E[V_1(k)] = O\Big(\frac{a((1-K^{-1})n)}{n^{\zeta_g}}\Big).
$$

Then, applying (B.17) to all K folds and using Jensen's inequality, we find that

$$E\Big[\sup_{\pi\in\Pi}\big|D_1(\pi)\big|\Big] = O\Big(\sqrt{\mathrm{VC}(\Pi)\frac{a((1-K^{-1})n)}{n^{1+\zeta_g}}}\Big). \tag{B.18}$$

We proceed to bound $D_2(\pi)$. Since for any integrable function $\tilde{m} : \mathcal{X}\times\mathcal{Z}\to\mathbb{R}$, $E[\tilde{m}(X,\alpha_1(Z)) - \tilde{m}(X,\alpha_0(Z)) - g(X,Z)\tilde{m}(X,Z)] = 0$,

$$\begin{aligned}E[\Delta\hat{\varphi}_i - \Delta\varphi_i - g(X_i,Z_i)(\hat{\varphi}^{(-k(i))}(X_i,\hat{p}^{(-k(i))}(X_i,Z_i);\hat{p}^{(-k(i))}) \\ - \varphi(X_i,p(X_i,Z_i);p))|\hat{p}^{(-k(i))}(\cdot),\hat{\varphi}^{(-k(i))}(\cdot;\hat{p}^{(-k(i))}(\cdot))] = 0.\end{aligned}$$

Thus, by a similar argument as before, we can show that

$$E\Big[\sup_{\pi\in\Pi}\big|D_2(\pi)\big|\Big] = O\Big(\sqrt{\mathrm{VC}(\Pi)\frac{a((1-K^{-1})n)}{n^{1+\zeta_\varphi}}}\Big). \tag{B.19}$$

It remains to bound $D_3(\pi)$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}|D_3(\pi)| \le &\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{\varphi}^{(-k(i))}(X_i,\hat{p}^{(-k(i))}(X_i,Z_i);\hat{p}^{(-k(i))}) - \varphi(X_i,p(X_i,Z_i);p))^2} \\ &\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{g}^{(-k(i))}(X_i,Z_i) - g(X_i,Z_i))^2}.\end{aligned}$$

Since this bound does not depend on π, we can apply the Cauchy-Schwarz inequality again to obtain

$$\begin{aligned}E\Big[\sup_{\pi\in\Pi}|D_3(\pi)|\Big] &\le \sqrt{E[(\hat{\varphi}^{(-k(i))}(X_i,\hat{p}^{(-k(i))}(X_i,Z_i);\hat{p}^{(-k(i))}) - \varphi(X_i,p(X_i,Z_i);p))^2]} \\ &\quad\cdot\sqrt{E[(\hat{g}^{(-k(i))}(X_i,Z_i) - g(X_i,Z_i))^2]} \\ &= O\Big(\frac{a((1-K^{-1})n)}{\sqrt{n}}\Big).\end{aligned} \tag{B.20}$$

Combining (B.18)–(B.20), we have

$$\sqrt{n}E[\sup\{|\hat{A}_n(\pi) - \tilde{A}_n(\pi)| : \pi\in\Pi\}] = O\Big(1 + \sqrt{\mathrm{VC}(\Pi)\frac{a((1-K^{-1})n)}{n^{\min\{\zeta_g,\zeta_\varphi\}}}}\Big),$$

which can be viewed as a counterpart of Lemma 4 of Athey and Wager (2021). The

desired result follows from the proof of Theorem 1 of Athey and Wager (2021). □

B.7 Proof of Auxiliary Lemmas

Proof of Lemma B.1. Setting $\Xi_i = n^{-1}(\tilde{b}^k(\tilde{X}_i)\tilde{b}^k(\tilde{X}_i)^\top - I_k)$ and noting that

$$\max_{1\le i\le n} \|\Xi_i\| \le n^{-1}(\zeta_k^2\lambda_k^2 + 1),$$
$$\max\left\{\left\|\sum_{i=1}^n E[\Xi_i\Xi_i^\top]\right\|, \left\|\sum_{i=1}^n E[\Xi_i^\top\Xi_i]\right\|\right\} \le n^{-1}(\zeta_k^2\lambda_k^2 + 1),$$

Then, the desired result follows from Theorem 4.1 of Chen and Christensen (2015). □

Proof of Lemma B.2. Fix $\tilde{x} \in \tilde{\mathcal{X}}$ and $\delta \in (0, 1]$. By rotational invariance, we have

$$\hat{p}(\tilde{x}) - \tilde{p}(\tilde{x}) = \tilde{b}^k(\tilde{x})^\top(\tilde{B}^\top\tilde{B}/n)^-\tilde{B}^\top e/n,$$

where $e = (\epsilon_1, \ldots, \epsilon_n)^\top$. Define $G_{i,n}(\tilde{x}) = \tilde{b}^k(\tilde{x})^\top(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{X}_i)1_{\mathcal{A}_n}$. Then $(\hat{p}(\tilde{x}) - \tilde{p}(\tilde{x}))1_{\mathcal{A}_n} = \frac{1}{n}\sum_{i=1}^n G_{i,n}(\tilde{x})\epsilon_i$. Note that $\|(\tilde{B}^\top\tilde{B}/n)^{-1}\| \le 2$ on $\mathcal{A}_n$. Then, for $q \ge 2$,

$$\begin{aligned}
&\sum_{i=1}^n E[(n^{-1}G_{i,n}(\tilde{x}))^2|\epsilon_i|^q|\tilde{X}_1^n] \\
&= \frac{1}{n^2}\sum_{i=1}^n E[|\epsilon_i|^q|\tilde{X}_i]\tilde{b}^k(\tilde{x})^\top(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{X}_i)\tilde{b}^k(\tilde{X}_i)^\top 1_{\mathcal{A}_n}(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{x}) \\
&\le \sup_{\tilde{x}} E[|\epsilon|^q|\tilde{X} = \tilde{x}]\frac{1}{n^2}\sum_{i=1}^n \tilde{b}^k(\tilde{x})^\top(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{X}_i)\tilde{b}^k(\tilde{X}_i)^\top 1_{\mathcal{A}_n}(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{x}) \\
&\le \frac{q!}{2}\nu^2\tilde{c}^{q-2}\frac{1}{n}\tilde{b}^k(\tilde{x})^\top(\tilde{B}^\top\tilde{B}/n)^-(\tilde{B}^\top\tilde{B}/n)1_{\mathcal{A}_n}(\tilde{B}^\top\tilde{B}/n)^-\tilde{b}^k(\tilde{x}) \\
&\le \frac{q!}{2}\nu^2\tilde{c}^{q-2}\frac{2\zeta_k^2\lambda_k^2}{n}.
\end{aligned}$$

Moreover, for $q \geq 3$,

$$\begin{aligned}
&\sum_{i=1}^{n} E[|n^{-1}G_{i,n}(\tilde{x})\epsilon_i|^q|\tilde{X}_1^n] \\
&\leq n^{-(q-2)} \sup_{\tilde{x}} |\tilde{b}^k(\tilde{x})^\top (\tilde{B}^\top \tilde{B}/n)^- 1_{\mathcal{A}_n} \tilde{b}^k(\tilde{x})|^{q-2} \sum_{i=1}^{n} E[(n^{-1}G_{i,n}(\tilde{x}))^2 |\epsilon_i|^q |\tilde{X}_1^n] \\
&\leq \Big(\frac{2\zeta_k^2\lambda_k^2}{n}\Big)^{q-2} \frac{q!}{2}\nu^2 \tilde{c}^{q-2} \frac{2\zeta_k^2\lambda_k^2}{n} \\
&= \frac{q!}{2}\nu^2 \frac{2\zeta_k^2\lambda_k^2}{n} \Big(\tilde{c}\frac{2\zeta_k^2\lambda_k^2}{n}\Big)^{q-2}.
\end{aligned}$$

Hence, the conditions of Theorem 2.10 of Boucheron et al. (2013) hold with $v = \nu^2 \frac{2\zeta_k^2\lambda_k^2}{n}$ and $c = \tilde{c}\frac{2\zeta_k^2\lambda_k^2}{n}$. Since v and c do not depend on $\tilde{x}$, by their Corollary 2.11,

$$\begin{aligned}
\Pr(1_{\mathcal{A}_n}\|\hat{p} - \tilde{p}\|_\infty \geq \delta) &= \Pr\Big(\sup_{\tilde{x}} \Big|\frac{1}{n}\sum_{i=1}^{n} G_{i,n}(\tilde{x})\epsilon_i\Big| \geq \delta\Big) \\
&\leq \exp\Big(-\frac{n\delta^2}{4\zeta_k^2\lambda_k^2(\nu^2 + \tilde{c}\delta)}\Big) \\
&\leq \exp\Big(-\frac{C_6 n\delta^2}{\zeta_k^2\lambda_k^2}\Big)
\end{aligned}$$

for some finite positive constant C_6. □

Proof of Lemma B.3. First, for any $h \in B_k$,

$$\begin{aligned}
\|\tilde{p} - p\|_\infty &= \|\tilde{p} - h + h - p\|_\infty \\
&= \|P_{k,n}(p - h) + h - p\|_\infty \\
&\leq \|P_{k,n}(p - h)\|_\infty + \|p - h\|_\infty \\
&\leq (1 + \|P_{k,n}\|_\infty)\|p - h\|_\infty.
\end{aligned}$$

Taking the infimum over $h \in B_k$ yields the first result. Second, take any $h \in L^\infty(\tilde{X})$ with $\|h\|_\infty \neq 0$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|1_{\mathcal{A}_n} P_{k,n} h(\tilde{x})| &\leq \|\tilde{b}^k(\tilde{x})\| \|1_{\mathcal{A}_n}(\tilde{B}^\top \tilde{B}/n)^- \tilde{B}^\top \mathbf{H}/n\| \\
&\leq \zeta_k \lambda_k \|1_{\mathcal{A}_n}(\tilde{B}^\top \tilde{B}/n)^- \tilde{B}^\top \mathbf{H}/n\|
\end{aligned}$$

uniformly over $\tilde{x}$. On $\mathcal{A}_n$, $\|\tilde{B}^\top\tilde{B}/n - I_k\| \leq \frac{1}{2}$ and thus $\lambda_{\min}(\tilde{B}^\top\tilde{B}/n) \geq \frac{1}{2}$. Then,

$$\begin{aligned}\|1_{\mathcal{A}_n}(\tilde{B}^\top\tilde{B}/n)^-\tilde{B}^\top\mathbf{H}/n\|^2 &= 1_{\mathcal{A}_n}(\mathbf{H}^\top\tilde{B}/n(\tilde{B}^\top\tilde{B}/n)^{-1}(\tilde{B}^\top\tilde{B}/n)^{-1}\tilde{B}^\top\mathbf{H}/n \\ &\leq 2\mathbf{H}^\top\tilde{B}/n(\tilde{B}^\top\tilde{B}/n)^{-1}\tilde{B}^\top\mathbf{H}/n \\ &\leq 2\|h\|^2 \\ &\leq 2\|h\|_\infty^2,\end{aligned}$$

where the second-last inequality follows from $1_{\mathcal{A}_n}\tilde{B}(\tilde{B}^\top\tilde{B})^{-1}\tilde{B}^\top$ being idempotent. It follows that $\|1_{\mathcal{A}_n}P_{k,n}h\|_\infty/\|h\|_\infty \leq \sqrt{2}\zeta_k\lambda_k$ uniformly in h. Taking the sup over h yields the second result. □

Proof of Lemma B.4. For any p and φ, we have

$$E[m(X,Z;p,\varphi)] = E[\pi(X,Z)g(X,Z)\varphi(X,p(X,Z);p)].$$

The chain rule gives

$$\begin{aligned}\frac{\partial}{\partial\tau}E[m(X,Z;p_\tau,\varphi_\tau)] = &\frac{\partial}{\partial\tau}E[\pi(X,Z)g(X,Z)\varphi_\tau(X,p_0(X,Z);p_0)] \\ &+ \frac{\partial}{\partial\tau}E[\pi(X,Z)g(X,Z)\varphi_0(X,p_\tau(X,Z);p_0)] \\ &+ \frac{\partial}{\partial\tau}E[\pi(X,Z)g(X,Z)\varphi_0(X,p_0(X,Z);p_\tau)]. \qquad \text{(B.21)}\end{aligned}$$

For the first term on the right-hand side of (B.21), note that

$$\begin{aligned}&\frac{\partial}{\partial\tau}E[\pi(X,Z)g(X,Z)\varphi_\tau(X,p_0(X,Z);p_0)] \\ =\ &\frac{\partial}{\partial\tau}E_\tau[\pi(X,Z)g(X,Z)\varphi_\tau(X,p_0(X,Z);p_0)] \\ &- \frac{\partial}{\partial\tau}E_\tau[\pi(X,Z)g(X,Z)\varphi_0(X,p_0(X,Z);p_0)] \\ =\ &\frac{\partial}{\partial\tau}E_\tau[\pi(X,Z)g(X,Z)(Y-\varphi_0(X,p_0(X,Z);p_0))] \\ =\ &E_H[\pi(X,Z)g(X,Z)(Y-\varphi_0(X,p_0(X,Z);p_0))].\end{aligned}$$

The second and third terms on the right-hand side of (B.21) reflect the two contributions of the estimation error of p. Note that for any τ,

$$E[\pi(X,Z)g(X,Z)(Y-\varphi_0(X,p_\tau(X,Z);p_\tau))] = 0.$$

Differentiating with respect to τ and evaluating the result at $\tau = 0$, we find

$$\frac{\partial}{\partial \tau} E[\pi(X,Z)g(X,Z)\varphi_0(X, p_\tau(X,Z); p_0)] + \frac{\partial}{\partial \tau} E[\pi(X,Z)g(X,Z)\varphi_0(X, p_0(X,Z); p_\tau)] = 0.$$

Putting everything together yields the desired result. □

B.8 Additional Tables and Figures

Table B.2: Sample Averages for the Treatment and Control Groups

	Upper secondary or higher (treatment group) $N = 841$	Less than upper secondary (control group) $N = 1263$
Log hourly wages	8.018	7.209
Years of education	13.128	5.585
Distance to school (km)	1.529	1.565
Distance to health post (km)	0.331	0.361
Fees per continuing student (1000 Rupiah)	3.464	3.992
Age	35.668	36.766
Religion Protestant	0.037	0.008
Catholic	0.023	0.007
Other	0.075	0.039
Muslim	0.866	0.946
Father uneducated	0.189	0.254
elementary	0.325	0.251
secondary and higher	0.275	0.026
missing	0.212	0.468
Mother uneducated	0.182	0.203
elementary	0.301	0.173
secondary and higher	0.138	0.011
missing	0.379	0.612
Rural household	0.483	0.644
North Sumatra	0.034	0.045
West Sumatra	0.023	0.023
South Sumatra	0.069	0.033
Lampung	0.013	0.028
Jakarta	0	0
Central Java	0.102	0.216
Yogyakarta	0.121	0.077
East Java	0.152	0.201
Bali	0.081	0.035
West Nussa Tengara	0.084	0.053
South Kalimanthan	0.040	0.033
South Sulawesi	0.040	0.019

Appendix C

Appendix for Chapter 3

C.1 Proofs of Theorems 3.1 and 3.2

Proof of Theorem 3.1. For any subsequence of $\{n\}$, we modify the definition of I_1 and I_2 accordingly. It suffices to show that for any subsequence $\{b_n\}$ of $\{n\}$ and any $\{P_{b_n} \in \mathcal{P}_0\}$, there exists a further subsequence $\{a_n\}$ of $\{b_n\}$ such that $\lim_{n\to\infty} E_{P_{a_n}}[\varphi_{a_n}(\alpha)] = \alpha$ for $\varphi_n = \varphi_n^{\text{2-sided}}(\alpha)$ or $\varphi_n = \varphi_n^{\text{1-sided}}(\alpha)$. Fix $\ell \in \{1,2\}$. By the completeness of the real line, there is always a subsequence $\{a_n\}$ of $\{b_n\}$ such that $\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, , P_{a_n})) \xrightarrow{p} \sigma_\infty \in [0,\infty)$. By Lemmas C.3 and C.4,

$$
\begin{aligned}
&\sqrt{a_n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_{a_n}}) \\
&= o_p(1) + (a_n/2)^{-1/2} \sum_{i\in I_\ell} (\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})}(Y_i|X_i; p_{a_n,y|x}) - E_{P_{a_n}}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})}(Y|X; p_{a_n,y|x})]) \\
&= o_p(1) + \sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot (Z_\ell + o_p(1)) \\
&= \sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot Z_\ell + o_p(1). \qquad \text{(C.1)}
\end{aligned}
$$

We consider two cases.

Case 1: $\sigma_\infty = 0$. By Lemma C.5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = \sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) + o_p(1) = o_p(1)$. Then, by (C.1), $\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} = o_p(1) \cdot O_p(1) + o_p(1) = o_p(1)$. By Condition 3.1(a), we have

$$
\hat{T}_{I_\ell} = \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}} = \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell}/\hat{\omega}_{I_\ell} + U_\ell}{\sqrt{(\hat{\sigma}_{I_\ell}(\hat{\theta}_{I_{-\ell}})/\hat{\omega}_{I_\ell})^2 + 1}} = \frac{o_p(1) + U_\ell}{\sqrt{o_p(1)+1}} = U_\ell + o_p(1).
$$

Case 2: $\sigma_\infty > 0$. By Lemma C.5,

$$
\frac{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}})}{\sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n}))} = 1 + \frac{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) - \sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n}))}{\sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n}))} = 1 + o_p(1).
$$

Then, by (C.1) and Condition 3.1(b),

$$\begin{aligned}
\hat{T}_{I_\ell} &= \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}} \\
&= \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell}/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) + \hat{\omega}_{I_\ell}/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}})/\sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) + (\hat{\omega}_{I_\ell}/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})))^2}} \\
&= \frac{Z_\ell + o_p(1)}{\sqrt{1 + o_p(1)}} \\
&= Z_\ell + o_p(1).
\end{aligned}$$

Now, we add ℓ-superscripts to a_n and redefine $\{a_n\} = \cup_{\ell=1}^2 \{a_n^\ell\}$. Then, we can conclude that in all cases $\hat{T}_{a_n} \xrightarrow{d} N(0,1)$ and the desired result follows. □

Proof of Theorem 3.2. Let $\{b_n\}$ be a subsequence of $\{n\}$ such that

$$\lim_{n\to\infty} E_{P_{b_n}}[\varphi_{b_n}(\alpha)] = \liminf_{n\to\infty} E_{P_n}[\varphi_n(\alpha)]$$

for $\varphi_n = \varphi_n^{\text{2-sided}}$ or $\varphi_n = \varphi_n^{\text{1-sided}}$. Fix $\ell \in \{1,2\}$. We focus on the subsequence $\{a_n\}$ of $\{b_n\}$ defined in the proof of Theorem 3.1 and consider two cases.
Case 1: $\sigma_\infty = 0$. By Lemma C.5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = o_p(1)$. Then, by (C.1) and Condition 3.1(a),

$$\begin{aligned}
\hat{T}_{I_\ell} &= \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}} \\
&= \frac{\sqrt{a_n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_{a_n}})/\hat{\omega}_{I_\ell} + U_\ell + \sqrt{a_n/2}\,\mathrm{QLR}_{P_{a_n}}/\hat{\omega}_{I_\ell}}{\sqrt{(\hat{\sigma}_{I_\ell}(\hat{\theta}_{I_{-\ell}})/\hat{\omega}_{I_\ell})^2 + 1}} \\
&= \frac{o_p(1) + U_\ell + h/(\sqrt{2}\omega_\infty)}{\sqrt{o_p(1) + 1}} \\
&= U_\ell + h/(\sqrt{2}\omega_\infty) + o_p(1).
\end{aligned}$$

Case 2: $\sigma_\infty > 0$. By Lemma C.5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}})/\sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) = 1 + o_p(1)$. Then, by

(C.1) and Condition 3.1(b),

$$\begin{aligned}
\hat{T}_{I_\ell} &= \frac{\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}} \\
&= \frac{\sqrt{a_n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_{a_n}})/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) + \hat{\omega}_{I_\ell}/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}})/\sigma^2_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) + (\hat{\omega}_{I_\ell}/\sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})))^2}} \\
&\quad + \frac{\sqrt{a_n/2}\,\mathrm{QLR}_{P_{a_n}}}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}} \\
&= \frac{Z_\ell + o_p(1)}{\sqrt{1 + o_p(1)}} + \frac{h/\sqrt{2}}{\sqrt{\sigma^2_\infty + o_p(1)}} + o_p(1) \\
&= Z_\ell + h/(\sqrt{2}\sigma_\infty) + o_p(1).
\end{aligned}$$

Now, we add ℓ-superscripts to a_n and redefine $\{a_n\} = \cup_{\ell=1}^2 \{a_n^\ell\}$. Let Z be an auxiliary random variable drawn from $N(0,1)$. Then, we can conclude that in all cases,

$$\begin{aligned}
\lim_{n\to\infty} E_{P_{a_n}}[\phi^{\text{2-sided}}_{a_n}(\alpha)] &= \lim_{n\to\infty} \Pr_{P_{a_n}}(|\hat{T}_{a_n}| > z_{1-\alpha/2}) \\
&\geq \Pr(|Z + h/(\omega_\infty \vee \sigma_\infty)| > z_{1-\alpha/2}) \\
&= 1 - \Phi(z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)) + \Phi(-z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)), \\
\lim_{n\to\infty} E_{P_{a_n}}[\phi^{\text{1-sided}}_{a_n}(\alpha)] &= \lim_{n\to\infty} \Pr_{P_{a_n}}(\hat{T}_{a_n} > z_{1-\alpha}) \\
&\geq \Pr(Z + h/(\omega_\infty \vee \sigma_\infty) > z_{1-\alpha}) \\
&= 1 - \Phi(z_{1-\alpha} - h/(\omega_\infty \vee \sigma_\infty)).
\end{aligned}$$

Therefore, the desired result follows. □

C.2 Auxiliary Lemmas and Their Proofs

Lemma C.1. *Suppose that Assumption 3.2 holds. Then, for each $s \in \{0,1\}$, (i) $L(x, \theta_s, p_{y|x})$ is differentiable with respect to θ_s; (ii) for any $\theta_s \in \Theta_s$ and $p_{y|x}, \tilde{p}_{y|x} \in \mathcal{H}$, $p_{y|x} \mapsto L(x, \theta_s, p_{y|x})$ is directionally differentiable at $p_{y|x}$, and the directional derivative at $p_{y|x}$ in the direction $\tilde{p}_{y|x}(y|x) - p_{y|x}(y|x)$ is given by*

$$D(x, \theta_s, p_{y|x}, \tilde{p}_{y|x} - p_{y|x}) = \sum_{y \in \mathcal{Y}} (\tilde{p}_{y|x}(y|x) - p_{y|x}(y|x)) \ln q^*_{\theta_s, y|x}(y|x; p_{y|x}).$$

Proof of Lemma C.1. For part (i), note that by definition,

$$L(x, \theta_s, p_{y|x}) = \max_{q \in \Delta} \sum_{y \in \mathcal{Y}} p_{y|x}(y|x) \ln q(y) \tag{C.2}$$
$$s.t. \quad \nu_{\theta_s}(A|x) \leq \sum_{y \in A} q(y), \ A \in \mathcal{C}.$$

Hence, the desired result follows from Theorem 3.1(i) of Kaido and Molinari (2024) by replacing $p_{0,y|x}$ with $p_{y|x}$. Part (ii) follows from Theorem 4.1 of Fiacco and Ishizuka (1990) by noting that (C.2) has a unique solution $q^*_{\theta_s,y|x}(\cdot|x; p_{y|x})$. □

Lemma C.2. *Suppose that Assumptions 3.1–3.4 hold. Then, for each $s, \ell \in \{1, 2\}$, if $\hat{\theta}_{s,I_\ell}$ approximately solves the first-order conditions: $\frac{2}{n} \sum_{i \in I_\ell} m(X_i, \hat{\theta}_{s,I_\ell}, \hat{p}_{I_\ell,y|x}) = o_p(n^{-d_p})$, then $d(\hat{\theta}_{s,I_\ell}, \Theta^*_s(P_0)) = O_p(n^{-d_p})$.*

Proof of Lemma C.2. We omit s-subscripts and the sample-splitting feature for readability. Define $Q_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^n m(X_i, \theta, \hat{p}_{n,y|x})$. For each $\epsilon > 0$, define the ϵ-expansion of $\Theta^*(P_0)$ in Θ by $\Theta^\epsilon(P_0) \equiv \{\theta \in \Theta : d(\theta, \Theta^*(P_0)) \leq \epsilon\}$. For the desired result, it suffices to show that for some $\epsilon_n = O(n^{-d_p})$ with $\epsilon_n > 0$, $\hat{\theta}_n \in \Theta^{\epsilon_n}(P_0)$ with probability approaching 1. On one hand, we can write

$$\begin{aligned} n^{d_p} Q_n(\theta) = n^{d_p} E_{P_0}[m(X, \theta, p_{0,y|x})] + n^{d_p - 1/2} \mathbb{G}_n(m(\cdot, \theta, \hat{p}_{n,y|x})) \\ + n^{d_p}(E_{P_0}[m(X, \theta, \hat{p}_{n,y|x})] - E_{P_0}[m(X, \theta, p_{0,y|x})]). \end{aligned}$$

By Assumption 3.3(b), $\mathbb{G}_n(m(\cdot, \theta, \hat{p}_{n,y|x})) = O_p(1)$. By Assumptions 3.3(c) and 3.4, $E_{P_0}[m(X, \theta, \hat{p}_{n,y|x})] - E_{P_0}[m(X, \theta, p_{0,y|x})] = O_p(n^{-d_p})$. Hence, by Assumption 3.3(a),

$$\|n^{d_p} Q_n(\theta)\| = C \cdot n^{d_p}(d(\theta, \Theta^*(P_0)) \wedge \delta) + O_p(1).$$

Namely, for any $\varepsilon > 0$, there exist $M > 0$ and $n_\varepsilon > 0$ such that for all $n \geq n_\varepsilon$,

$$P(\|n^{d_p} Q_n(\theta)\| - C \cdot n^{d_p}(d(\theta, \Theta^*(P_0)) \wedge \delta) \geq -M) \geq 1 - \varepsilon.$$

Further, there exists $n_\delta > n_\varepsilon > 0$ such that for all $n \geq n_\delta$, $\frac{1}{2} C \cdot n^{d_p} \delta \geq M$. Also note that for any $\theta \in \Theta$ satisfying $d(\theta, \Theta^*(P_0)) \geq \frac{2M}{n^{d_p} C}$, $\frac{1}{2} C \cdot n^{d_p} d(\theta, \Theta^*(P_0)) \geq M$. It follows that for all $n \geq n_\delta$,

$$P\Big(\|n^{d_p} Q_n(\theta)\| \geq \frac{1}{2} C \cdot n^{d_p}(d(\theta, \Theta^*(P_0)) \wedge \delta)\Big) \geq 1 - \varepsilon$$

uniformly in $\{\theta \in \Theta : d(\theta, \Theta^*(P_0)) \geq \frac{2M}{n^{d_p}C}\}$. On the other hand, $\|n^{d_p}Q_n(\hat{\theta}_n)\| = o_p(1)$. Hence, for any $\varepsilon > 0$, there exists $n'_\varepsilon > 0$ such that for all $n \geq n'_\varepsilon$, $P(\|n^{d_p}Q_n(\hat{\theta}_n)\| \leq M) \geq 1-\varepsilon$. Let $\epsilon_n = \frac{2M}{n^{d_p}C}$ and $\bar{n} = n_\delta \vee n'_\varepsilon$. We can conclude that for all $n \geq \bar{n}$, with probability at least $1-2\varepsilon$, $\inf_{\theta \in \Theta \setminus \Theta^{\epsilon_n}(P_0)} \|n^{d_p}Q_n(\theta)\| \geq \frac{1}{2}C \cdot n^{d_p}(\epsilon_n \wedge \delta) = \frac{1}{2}C \cdot n^{d_p}\epsilon_n = M$ and $\|n^{d_p}Q_n(\hat{\theta}_n)\| \leq M$. Therefore, $\hat{\theta}_n \in \Theta^{\epsilon_n}(P_0)$ with probability approaching 1. □

Lemma C.3. *Suppose Assumptions 3.1–3.5 hold. Then for each $s, \ell \in \{1, 2\}$,*

$$(n/2)^{-1/2} \sum_{i \in I_\ell} L(X_i, \hat{\theta}_{s,I_{-\ell}}, \hat{p}_{I_\ell, y|x}) = (n/2)^{-1/2} \sum_{i \in I_\ell} \ln q^*_{\theta^*(\hat{\theta}_{s,I_{-\ell}}, P_0), y|x}(Y_i|X_i; p_{0,y|x}) + o_p(1).$$

Proof of Lemma C.3. We omit s-subscripts and the sample-splitting feature for readability. Write

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n L(X_i, \hat{\theta}_n, \hat{p}_{n,y|x}) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \ln q^*_{\theta^*(\hat{\theta}_n, P_0), y|x}(Y_i|X_i; p_{0,y|x}) \\
&= \mathbb{G}_n(L(\cdot, \hat{\theta}_n, \hat{p}_{n,y|x}) - L(\cdot, \theta^*(\hat{\theta}_n, P_0), p_{0,y|x})) \\
&\quad + \sqrt{n}(E_{P_0}[L(X, \hat{\theta}_n, \hat{p}_{n,y|x})] - E_{P_0}[L(X, \theta^*(\hat{\theta}_n, P_0), p_{0,y|x})]) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n (L(X_i, \theta^*(\hat{\theta}_n, P_0), p_{0,y|x})) - \ln q^*_{\theta^*(\hat{\theta}_n, P_0), y|x}(Y_i|X_i; p_{0,y|x})).
\end{aligned}$$

We examine each term on the right-hand side. First, by the mean value theorem and Assumption 3.5(a), for any $\theta, \tilde{\theta} \in \Theta$ and $p_{y|x}, \tilde{p}_{y|x} \in \mathcal{H}$,

$$|L(x, \theta, p_{y|x}) - L(x, \tilde{\theta}, \tilde{p}_{y|x})| \leq B(x)(\|\theta - \tilde{\theta}\| + \|p_{y|x} - \tilde{p}_{y|x}\|_{\mathcal{H}}).$$

Hence, by Assumption 3.5(b), we can apply Theorem 3 of Chen et al. (2003) to show that the empirical process $\mathbb{G}_n(L(\cdot, \theta_s, p_{y|x}))$ indexed by θ_s and $p_{y|x}$ is stochastically equicontinuous: for all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,

$$\sup_{\|\tilde{\theta}_s - \theta_s\| + \|\tilde{p}_{y|x} - p_{y|x}\|_{\mathcal{H}} \leq \delta_n} |\mathbb{G}_n(L(\cdot, \tilde{\theta}_s, \tilde{p}_{y|x}) - L(\cdot, \theta_s, p_{y|x}))| = o_p(1).$$

By Lemma C.2, $\|\hat{\theta}_n - \theta^*(\hat{\theta}_n, P_0)\| = O_p(n^{-d_p})$. Hence, for all $\delta_n = o(1)$, $\|\hat{\theta}_n - \theta^*(\hat{\theta}_n, P_0)\| \leq \delta_n$ with probability approaching 1. Similarly, by Assumption 3.4, $\|\hat{p}_{n,y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta_n$ with probability approaching 1. Therefore,

$$\mathbb{G}_n(L(\cdot, \hat{\theta}_n, \hat{p}_{n,y|x}) - L(\cdot, \theta^*(\hat{\theta}_n, P_0), p_{0,y|x})) = o_p(1).$$

Second, we can write

$$\begin{aligned}
&\sqrt{n}(E_{P_0}[L(X,\hat{\theta}_n,\hat{p}_{n,y|x})] - E_{P_0}[L(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})]) \\
=&\sqrt{n}(\hat{\theta}_n - \theta^*(\hat{\theta}_n,P_0))' E_{P_0}[m(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})] \\
&+ \sqrt{n}E_{P_0}[D(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x},\hat{p}_{n,y|x}-p_{0,y|x})] + \sqrt{n}E_{P_0}[r(X,\hat{\theta}_n,\hat{p}_{n,y|x})],
\end{aligned}$$

where

$$\begin{aligned}
r(x,\theta,p_{y|x}) \equiv L(x,\theta,p_{y|x}) - L(x,\theta^*(\theta,P_0),p_{0,y|x}) - (\theta-\theta^*(\theta,P_0))' m(x,\theta^*(\theta,P_0),p_{0,y|x}) \\
- D(x,\theta^*(\theta,P_0),p_{0,y|x},p_{y|x}-p_{0,y|x}).
\end{aligned}$$

By the first-order conditions for $\Theta^*(P_0)$,

$$E_{P_0}[m(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})] = 0.$$

By Jensen's inequality, Lemma C.2, and Assumptions 3.4 and 3.5(a),

$$\begin{aligned}
\sqrt{n}|E_{P_0}[r(X,\hat{\theta}_n,\hat{p}_{n,y|x})]| &\leq \sqrt{n}E_{P_0}[|r(X,\hat{\theta}_n,\hat{p}_{n,y|x})|] \\
&\leq E[B(X)]\sqrt{n}(\|\hat{\theta}_n - \theta^*(\hat{\theta}_n,P_0)\|^2 + \|\hat{p}_{n,y|x}-p_{0,y|x}\|_{\mathcal{H}}^2) = o_p(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(E_{P_0}[L(X,\hat{\theta}_n,\hat{p}_{n,y|x})] - E_{P_0}[L(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})]) \\
=&\sqrt{n}E_{P_0}[D(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x},\hat{p}_{n,y|x}-p_{0,y|x})] + o_p(1).
\end{aligned}$$

Third, we can write

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (L(X_i,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})) - \ln q^*_{\theta^*(\hat{\theta}_n,P_0),y|x}(Y_i|X_i;p_{0,y|x})) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha(Y_i,X_i),$$

where $\alpha(y,x)$ is defined in Assumption 3.5(c). Putting everything together, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n L(X_i,\hat{\theta}_n,\hat{p}_{n,y|x}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \ln q^*_{\theta^*(\hat{\theta}_n,P_0),y|x}(Y_i|X_i;p_{0,y|x}) \\
=&\sqrt{n}E_{P_0}[D(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x},\hat{p}_{n,y|x}-p_{0,y|x})] - \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha(Y_i,X_i) + o_p(1),
\end{aligned}$$

and the desired result follows from Assumption 3.5(c). □

Lemma C.4. *Suppose Assumptions 3.1–3.6 hold. Then for each $\ell \in \{1,2\}$ and any sequence $\{P_n \in \mathcal{P}\}$,*

$$(n/2)^{-1/2} \sum_{i \in I_\ell} \frac{\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y_i|X_i; p_{n,y|x}) - E_{P_n}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y|X; p_{n,y|x})]}{\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n))} = Z_\ell + o_p(1),$$

where $Z_\ell \sim N(0,1)$ and Z_1 and Z_2 are independent.

Proof of Lemma C.4. Fix $\ell \in \{1,2\}$. Define the triangular array

$$Z_{ni} = \frac{\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y_i|X_i; p_{n,y|x}) - E_{P_n}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y|X; p_{n,y|x})]}{\sqrt{n/2}\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n))}, \ n \in \mathbb{N}_+, i \in I_\ell,$$

so that we can write

$$(n/2)^{-1/2} \sum_{i \in I_\ell} \frac{\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y_i|X_i; p_{n,y|x}) - E_{P_n}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y|X; p_{n,y|x})]}{\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n))} = \sum_{i \in I_\ell} Z_{ni}.$$

We verify the Lyapounov condition for $\{Z_{ni} : n \in \mathbb{N}_+, i \in I_\ell\}$. For any $\epsilon > 0$ and $n \in \mathbb{N}_+$,

$$\begin{aligned}
&\sum_{i \in I_\ell} E_{P_n}[|Z_{ni}|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}] \\
&= \frac{\sum_{i \in I_\ell} E_{P_n}[|\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y_i|X_i; p_{n,y|x}) - E_{P_n}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)}(Y|X; p_{n,y|x})]|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}]}{(\sqrt{n/2}\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)))^{2+\epsilon}} \\
&\leq \frac{(n/2)E_{P_n}[|D(Y,X)\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n))|^{2+\epsilon}]}{(\sqrt{n/2}\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)))^{2+\epsilon}} \\
&= (n/2)^{-\epsilon/2} E_{P_n}[|D(Y,X)|^{2+\epsilon}],
\end{aligned}$$

where the inequality follows from the independence between I_1 and I_2 and Assumption 3.6. Hence, there exist $M, \epsilon > 0$ such that for each $n \in \mathbb{N}_+$, $\sum_{i \in I_\ell} E_{P_n}[|Z_{ni}|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}] \leq (n/2)^{-\epsilon/2}M$. By the law of iterated expectations, the Lyapounov condition holds:

$$\sum_{i \in I_\ell} E_{P_n}[|Z_{ni}|^{2+\epsilon}] \leq (n/2)^{-\epsilon/2}M \to 0.$$

Then, we can apply Lyapounov's Central Limit Theorem to obtain $\sum_{i\in I_\ell} Z_{ni} \xrightarrow{d} N(0,1)$. By Skorohod's representation theorem and Lemma 9 of Chernozhukov et al. (2013b), if we enrich the original probability space $(\Omega, \mathcal{B}, P)$ by creating a new space as the product of $(\Omega, \mathcal{B}, P)$ and $([0,1], \mathcal{F}, \lambda)$, where $\mathcal{F}$ is the Borel sigma algebra on $[0,1]$ and λ is the Lebesgue measure, we have

$$\sum_{i\in I_\ell} Z_{ni} = Z_\ell + o_p(1),$$

where $Z_\ell \sim N(0,1)$ is independent of $\sum_{i\in I_\ell} Z_{ni}$. It follows that Z_1 and Z_2 are independent. □

Lemma C.5. *Suppose Assumptions 3.1–3.7 hold. Then for each $\ell \in \{1,2\}$,*

$$\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) - \sigma^2_{P_0}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_0)) = O_p(n^{-d_p}).$$

Proof of Lemma C.5. We omit the sample-splitting feature for readability. Define

$$\begin{aligned} V(\theta, p_{y|x}) &= E_{P_0}[\lambda^2_\theta(Y|X; p_{y|x})] - E_{P_0}[\lambda_\theta(Y|X; p_{y|x})]^2, \\ V_n(\theta, p_{y|x}) &= \frac{1}{n}\sum_{i=1}^n \lambda_\theta(Y_i|X_i; p_{y|x}) - \Big(\frac{1}{n}\sum_{i=1}^n \lambda_\theta(Y_i|X_i; p_{y|x})\Big)^2. \end{aligned}$$

We can write

$$\begin{aligned} (\hat{\sigma}_n(\hat{\theta}_n))^2 - \sigma^2_{P_0}(\theta^*(\hat{\theta}_n, P_0)) = (V_n(\hat{\theta}_n, \hat{p}_{n,y|x}) - V(\hat{\theta}_n, \hat{p}_{n,y|x})) \\ + (V(\hat{\theta}_n, \hat{p}_{n,y|x}) - V(\theta^*(\hat{\theta}_n, P_0), p_{0,y|x})). \end{aligned}$$

We examine each term on the right-hand side. First, by Assumption 3.7(b),

$$\begin{aligned} V_n(\hat{\theta}_n, \hat{p}_{n,y|x}) - V(\hat{\theta}_n, \hat{p}_{n,y|x}) &= n^{-1/2}\mathbb{G}_n(\lambda^2_{\hat{\theta}_n}(\cdot|\cdot; \hat{p}_{n,y|x})) - n^{-1/2}\mathbb{G}_n(\lambda_{\hat{\theta}_n}(\cdot|\cdot; \hat{p}_{n,y|x})) \\ &\quad \cdot (2E_{P_0}[\lambda_{\hat{\theta}_n}(Y|X; \hat{p}_{n,y|x})] + n^{-1/2}\mathbb{G}_n(\lambda_{\hat{\theta}_n}(\cdot|\cdot; \hat{p}_{n,y|x}))) \\ &= O_p(n^{-1/2}) - O_p(n^{-1/2}) \cdot (O_p(1) + O_p(n^{-1/2})) \\ &= O_p(n^{-1/2}). \end{aligned}$$

Second, by the triangle inequality and Jensen's inequality,

$$\begin{aligned}
&|V(\hat{\theta}_n, \hat{p}_{n,y|x}) - V(\theta^*(\hat{\theta}_n, P_0), p_{0,y|x})| \\
\leq\ &|E_{P_0}[\lambda^2_{\hat{\theta}_n}(Y|X; \hat{p}_{n,y|x})] - E_{P_0}[\lambda^2_{\theta^*(\hat{\theta}_n, P_0)}(Y|X; p_{0,y|x})]| \\
&+ 2|E_{P_0}[\lambda_{\hat{\theta}_n}(Y|X; \hat{p}_{n,y|x})] - E_{P_0}[\lambda_{\theta^*(\hat{\theta}_n, P_0)}(Y|X; p_{0,y|x})]| \\
&\cdot \Big(\sum_{s=1}^{2} \sup_{\theta_s \in \Theta_s, p_{y|x} \in \mathcal{H}} E_{P_0}[|\ln q^*_{\theta_s, y|x}(Y|X; p_{y|x})|] \Big).
\end{aligned}$$

By Lemma C.2 and Assumptions 3.4 and 3.7(c),

$$|E_{P_0}[\lambda^k_{\hat{\theta}_n}(Y|X; \hat{p}_{n,y|x})] - E_{P_0}[\lambda^k_{\theta^*(\hat{\theta}_n, P_0)}(Y|X; p_{0,y|x})]| = O_p(n^{-d_p}),\ \ k = 1, 2.$$

By Assumption 3.7(a), $\sup_{\theta_s \in \Theta_s, p_{y|x} \in \mathcal{H}} E_{P_0}[|\ln q^*_{\theta_s, y|x}(Y|X; p_{y|x})|] = O_p(1)$ for each $s \in \{1, 2\}$. It follows that

$$|V(\hat{\theta}_n, \hat{p}_{n,y|x}) - V(\theta^*(\hat{\theta}_n, P_0), p_{0,y|x})| \leq O_p(n^{-d_p}) + O_p(n^{-d_p}) \cdot O_p(1) = O_p(n^{-d_p}).$$

Putting everything together yields the desired result. □

Lemma C.6. *Suppose that Assumptions 3.1–3.7 hold. Then, for each $\ell \in \{1, 2\}$, $\hat{\omega}_{I_\ell}$ defined in (3.20) satisfies Condition 3.1.*

Proof of Lemma C.6. Fix $\ell \in \{1, 2\}$. To check Condition 3.1(a), note that by Lemma C.5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = \sigma^2_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n)) + [\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) - \sigma^2_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_n))] = O_p(n^{-d_\sigma}) + O_p(n^{-d_p}) = o_p((\ln n)^{-1})$. Hence, $\hat{\omega}_{I_\ell} \xrightarrow{p} 1$. To check Condition 3.1(b), note that by Lemma C.5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = \sigma^2_\infty + o_p(1)$. Hence, $\hat{\omega}_{I_\ell} \xrightarrow{p} 0$. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2021). Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6):2499–2529.

Barseghyan, L., Prince, J., and Teitelbaum, J. C. (2011). Are risk preferences stable across contexts? evidence from insurance data. *American Economic Review*, 101(2):591–631.

Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp Identification Regions in Models With Convex Moment Predictions. *Econometrica*, 79(6):1785–1821.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4):889–917.

Berry, S. T. and Compiani, G. (2022). An Instrumental Variable Approach to Dynamic Models. *The Review of Economic Studies*, 90(4):1724–1758.

Bertsimas, D. and McCord, C. (2018). Optimization over continuous and multidimensional decisions with observational data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bhattacharya, D. (2015). Nonparametric welfare analysis for discrete choice. *Econometrica*, 83(2):617–649.

Bhattacharya, D. (2018). Empirical welfare analysis for discrete choice: Some general results. *Quantitative Economics*, 9(2):571–615.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Björklund, A. and Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 69(1):42–49.

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

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in Semiparametric Binary Response Models. *The Review of Economic Studies*, 71(3):655–679.

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

Botosaru, I., Muris, C., and Pendakur, K. (2023). Identification of time-varying transformation models with fixed effects, with an application to unobserved heterogeneity in resource shares. *Journal of Econometrics*, 232(2):576–597.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bresnahan, T. F. and Reiss, P. C. (1990). Entry in monopoly market. *The Review of Economic Studies*, 57(4):531–553.

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039.

Byambadalai, U. (2022). Identification and inference for welfare gains without unconfoundedness. arXiv: 2207.04314.

Caballero, R. J. and Engel, E. M. R. A. (1999). Explaining investment dynamics in u.s. manufacturing: A generalized (s, s) approach. *Econometrica*, 67(4):783–826.

Cameron, S. V. and Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of american males. *Journal of Political Economy*, 106(2):262–333.

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, 78(1):377–394.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81.

Carneiro, P. and Lee, S. (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208.

Carneiro, P., Lokshin, M., and Umapathi, N. (2017). Average and marginal returns to upper secondary schooling in indonesia. *Journal of Applied Econometrics*, 32(1):16–36.

Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140(2):503–528.

Cattaneo, M. D., Jansson, M., and Nagasawa, K. (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica*, 88(5):2203–2219.

Chen, S. and Kaido, H. (2023). Robust tests of model incompleteness in the presence of nuisance parameters. arXiv: 2208.11281.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao.

Chen, X., Hong, H., and Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *Journal of Econometrics*, 141(1):109–140. Semiparametric methods in econometrics.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Chen, Y.-C. and Xie, H. (2022). Personalized subsidy rules. arXiv: 2202.13545.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013a). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

Chernozhukov, V., Fernández-Val, I., and Newey, W. K. (2019). Nonseparable multinomial choice models in cross-section and panel data. *Journal of Econometrics*, 211(1):104–116. Annals Issue in Honor of Jerry A. Hausman.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013b). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Chesher, A. and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3):959–989.

Chesher, A., Rosen, A. M., and Zhang, Y. (2024). Robust analysis of short panels. arXiv: 2401.06611.

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

Christensen, T., Moon, H. R., and Schorfheide, F. (2023). Optimal decision rules when payoffs are partially identified. arXiv:2204.11748.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Coate, S. and Conlin, M. (2004). A group rule-utilitarian approach to voter turnout: Theory and evidence. *American Economic Review*, 94(5):1476–1504.

Cohen, A. and Einav, L. (2007). Estimating risk preferences from deductible choice. *American Economic Review*, 97(3):745–788.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409.

Cui, Y. and Tchetgen Tchetgen, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533):162–173. PMID: 33994604.

D'Adamo, R. (2022). Orthogonal policy learning under ambiguity. arXiv: 2111.10904.

Davezies, L., D'Haultfoeuille, X., and Laage, L. (2024). Identification and estimation of average marginal effects in fixed effects logit models. arXiv: 2105.00879.

Dickstein, M. J. and Morales, E. (2018). What do Exporters Know?*. *The Quarterly Journal of Economics*, 133(4):1753–1801.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Eizenberg, A. (2014). Upstream innovation and product variety in the u.s. home pc market. *The Review of Economic Studies*, 81(3 (288)):1003–1045.

Fack, G., Grenet, J., and He, Y. (2019). Beyond truth-telling: Preference estimation with centralized school choice and college admissions. *American Economic Review*, 109(4):1486–1529.

Fafchamps, M. (1993). Sequential labor decisions under uncertainty: An estimable household model of west-african farmers. *Econometrica*, 61(5):1173–1197.

Fiacco, A. V. and Ishizuka, Y. (1990). Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, 27(1):215–235.

Francois, P., Rainer, I., and Trebbi, F. (2015). How is power shared in africa? *Econometrica*, 83(2):465–503.

Galichon, A. and Henry, M. (2011). Set Identification in Models with Multiple Equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

Gao, W. Y. and Li, M. (2024). Robust semiparametric estimation in panel multinomial choice models. arXiv: 2009.00085.

Gao, W. Y. and Wang, R. (2024). Identification of nonlinear dynamic panels under partial stationarity. arXiv: 2401.00264.

Goeree, M. S. (2008). Limited Information and Advertising in the U.S. Personal Computer Industry. *Econometrica*, 76(5):1017–1074.

Gu, J., Russell, T., and Stringham, T. (2024). Counterfactual identification and latent space enumeration in discrete outcome models. Available at SSRN 4188109.

Hahn, J. and Ridder, G. (2013). Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340.

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, 111(1).

Heckman, J., Tobias, J. L., and Vytlacil, E. (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 68(2):211–223.

Heckman, J. and Vytlacil, E. (2001a). Local instrumental variables. In Hsiao, C., Morimune, K., and Powell, J. L., editors, *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*. Cambridge University Press.

Heckman, J. J. and Vytlacil, E. (2001b). Policy-relevant treatment effects. *The American Economic Review*, 91(2):107–111.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4730–4734.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Hirano, K. and Porter, J. R. (2020). Chapter 4 - asymptotic analysis of statistical decision rules in econometrics. In Durlauf, S. N., Hansen, L. P., Heckman, J. J., and Matzkin, R. L., editors, *Handbook of Econometrics, Volume 7A*, pages 283–354. Elsevier.

Hoderlein, S. and White, H. (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168(2):300–314.

Honoré, B. E. and Kyriazidou, E. (2000). Estimation of tobit-type models with individual specific effects. *Econometric Reviews*, 19:341–66.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74(3):611–629.

Hsu, Y.-C. and Shi, X. (2017). Model-selection tests for conditional moment restriction models. *The Econometrics Journal*, 20(1):52–85.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Kaido, H. and Molinari, F. (2024). Information based inference in models with set-valued predictions and misspecification. arXiv:2401.11046.

Kaido, H., Molinari, F., and Stoye, J. (2022). Constraint qualifications in partial identification. *Econometric Theory*, 38(3):596–619.

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1243–1251. PMLR.

Kasy, M. (2016). Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies. *The Review of Economics and Statistics*, 98(1):111–131.

Kendall, C., Nannicini, T., and Trebbi, F. (2015). How do voters respond to information? evidence from a randomized campaign. *American Economic Review*, 105(1):322–53.

Khan, S., Ouyang, F., and Tamer, E. (2021). Inference on semiparametric multinomial response models. *Quantitative Economics*, 12(3):743–777.

Khan, S., Ponomareva, M., and Tamer, E. (2016). Identification of panel data models with endogenous censoring. *Journal of Econometrics*, 194(1):57–75.

Khan, S., Ponomareva, M., and Tamer, E. (2023). Identification of dynamic binary response models. *Journal of Econometrics*, 237(1):105515.

Kitagawa, T. and Tetenov, A. (2018a). Supplement to 'who should be treated? empirical welfare maximization methods for treatment choice'. *Econometrica Supplemental Material*, 86(2).

Kitagawa, T. and Tetenov, A. (2018b). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Kédagni, D. and Mourifié, I. (2020). Generalized Instrumental Inequalities: Testing the Instrumental Variable Independence Assumption. *Biometrika*, 107(3):661–675.

Li, T. (2009). Simulation based selection of competing structural econometric models. *Journal of Econometrics*, 148(2):114–123.

Liao, Z. and Shi, X. (2020). A nondegenerate Vuong test and post selection confidence intervals for semi/nonparametric models. *Quantitative Economics*, 11(3):983–1017.

Liu, L., Poirier, A., and Shiu, J.-L. (2024). Identification and estimation of partial effects in nonlinear semiparametric panel models. *Journal of Econometrics*, page 105860.

Mammen, E., Rothe, C., and Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132 – 1170.

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

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review*, 48(4):1393–1410.

Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848.

Miyauchi, Y. (2016). Structural estimation of pairwise stable networks with nonnegative externality. *Journal of Econometrics*, 195(2):224 – 235.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–98.

Molchanov, I. (2005). *Theory of Random Sets*. Probability and Its Applications. Springer London.

Molinari, F. (2020). Chapter 5 - microeconometrics with partial identification. In Durlauf, S. N., Hansen, L. P., Heckman, J. J., and Matzkin, R. L., editors, *Handbook of Econometrics, Volume 7A*, pages 355–486. Elsevier.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Nyarko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005.

Paap, R. and Franses, P. H. (2000). A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables. *Journal of Applied Econometrics*, 15(6):717–744.

Pakel, C. and Weidner, M. (2024). Bounds on average effects in discrete choice panel data models. arXiv: 2309.09299.

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

Palfrey, T. R. and Prisbrey, J. E. (1997). Anomalous behavior in public goods experiments: How much and why? *The American Economic Review*, 87(5):829–846.

Paulson, A. L., Townsend, R. M., and Karaivanov, A. (2006). Distinguishing limited liability from moral hazard in a model of entrepreneurship. *Journal of Political Economy*, 114(1):100–144.

Pu, H. and Zhang, B. (2021). Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):318–345.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. (2021). Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191. PMID: 33731969.

Rivers, D. and Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal*, 5(1):1–39.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Russell, T. M. (2020). Policy transforms and learning optimal policies. arXiv:2012.11046.

Sasaki, Y. and Ura, T. (2021). Estimation and inference for policy relevant treatment effects. *Journal of Econometrics*.

Sasaki, Y. and Ura, T. (2024). Welfare analysis via marginal treatment effects. *Econometric Theory*, page 1–24.

Schennach, S. M. and Wilhelm, D. (2017). A simple parametric model selection test. *Journal of the American Statistical Association*, 112(520):1663–1674.

Semenova, V. (2024). Aggregated intersection bounds and aggregated minimax values. arXiv: 2303.00982.

Sheng, S. (2020). A structural econometric analysis of network formation games through subnetworks. *Econometrica*, 88(5):1829–1858.

Shi, X. (2015a). Model selection tests for moment inequality models. *Journal of Econometrics*, 187(1):1–17.

Shi, X. (2015b). A nondegenerate vuong test. *Quantitative Economics*, 6(1):85–121.

Shi, X., Shum, M., and Song, W. (2018). Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. *Econometrica*, 86(2):737–761.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

Sun, L. (2024). Empirical welfare maximization with constraints. arXiv: 2103.15298.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165.

Tebaldi, P., Torgovitsky, A., and Yang, H. (2023). Nonparametric estimates of demand in the california health insurance exchange. *Econometrica*, 91(1):107–146.

Thornton, R. L. (2008). The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–63.

Torgovitsky, A. (2019). Nonparametric inference on state dependence in unemployment. *Econometrica*, 87(5):1475–1505.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press.

Yata, K. (2025). Optimal decision rules under partial identification. arXiv: 2111.04926.

CURRICULUM VITAE

Yan Liu

Education

Ph.D., Economics, Boston University, Boston MA, May 2025 (expected)

Dissertation Title: Essays on the Econometric Analysis of Counterfactual Policies and Incompleteness

Main advisor: Hiroaki Kaido

Dissertation Committee: Hiroaki Kaido, Iván Fernández-Val, and Jean-Jacques Forneron

M.A., Economics, Kyoto University, Kyoto, Japan, 2018

B.A., Economics, Peking University, Beijing, China, 2015

B.S., Statistics, Peking University, Beijing, China, 2015

Fields of Interest

Econometrics, Applied Microeconomics

Publications

"Asymptotic Properties of the Maximum Likelihood Estimator in Regime-Switching Models with Time-Varying Transition Probabilities," (with Chaojun, Li) *The Econometrics Journal*, (2023) 26(1): 67-87.

Working Papers

"Robust Counterfactual Analysis for Nonlinear Panel Data Models," October 2024.

"Policy Learning under Endogeneity Using Instrumental Variables," March 2024.

"Model Selection Tests for Incomplete Models," (with Hiroaki Kaido), October 2024.

Conference and Seminar Presentations

2025: University of Notre Dame, Osaka University, Kyoto University (Institute of Economic Research), Washington University in St. Louis, University of Sydney Business School, 2025 CES North American Conference (UMich)

2024: AMES2024-China (ZJU), IAAE 2024 Annual Conference (XMU), BU-BC Joint Workshop in Econometrics

2023: BU-BC Joint Workshop in Econometrics

2022: NASMES2022 (Miami), IAAE 2022 Annual Conference (KCL), ESAM2022 (virtual), SETA2022 (virtual), YES2022 (Yale), MEG 2022 Conference (MSU)

2019: AMES2019 (XMU)

2018: AMES2018 (Sogang University)

Fellowships and Awards

Best Second Year Paper Award, Department of Economics, Boston University, 2022

Dean's Fellowship, Boston University, 2019-2024

Foreign Student Scholarship, Nomura Foundation, 2018-2019

Outstanding Master's Thesis Award, Graduate School of Economics, Kyoto University, 2018

Asian Future Leaders Scholarship Program, Bai Xian Asia Institute, 2016-2018

Academic Excellence Award, Peking University, 2013

Leo KoGuan Scholarship, Peking University, 2013

Merit Student, Peking University, 2012

ICBC Scholarship, Peking University, 2012

Work Experience

Research Assistant to Professor Hiroaki Kaido, Boston University, Fall 2021, Fall 2022, Spring 2024

Research Assistant to Professor Pierre Perron, Boston University, Fall 2024, Spring 2025

Research Fellow (DC2), Japan Society for the Promotion of Science, 2019

Referee Experience

Annals of Statistics

Teaching Experience

Teaching Assistant, Advanced Econometrics 1 (Ph.D.), Boston University, Spring 2021, Spring 2022, Spring 2023

Teaching Assistant, Empirical Economics 1, Boston University, Fall 2021

Teaching Assistant, Elementary Mathematical Economics, Boston University, Fall 2020

Teaching Assistant, Advanced Microeconomics, Kyoto University, Spring 2019

Teaching Assistant, Advanced Econometrics, Kyoto University, Fall 2018

Teaching Assistant, Introduction to Financial Economics, Peking University, Fall 2014

Teaching Assistant, Principles of Economics, Peking University, Fall 2014

Department Service

Co-Organizer, BU Econometrics Reading Group, Fall 2021-Spring 2023

Languages

Chinese (native), English (fluent), Japanese (fluent)

Computer Skills: MATLAB, Python, R, STATA, LaTeX
Citizenship/Visa Status: China/F1