Model Selection Tests for Incomplete Models*

Hiroaki Kaido
Department of Economics
Boston University
hkaido@bu.edu

Yan Liu
Department of Economics
Boston University
yanliu@bu.edu

November 3, 2024

Abstract

This paper 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.

Keywords: Model selection, Likelihood-ratio tests, Incomplete models

*Financial support from NSF grant SES-2018498 is gratefully acknowledged. We are grateful to Francesca Molinari for her input on an early version of the paper. We thank participants of the BU Econometrics Reading Group for their comments.

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 paper 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 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.

[1] See, for example, 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.

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, 1996). 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 data-generating processes. 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 data-generating process, 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 paper 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 consider 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 $supp(A)$ and $supp(A|B)$, respectively.

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 $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$:

$$Y \in G(U|X;\theta), \ a.s. \tag{1}$$

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

$$Y = g(U|X;\theta). \tag{2}$$

The structure in (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 1 (Discrete Games): Consider a binary-response game with two players (e.g. firms). Each player may either choose $y^{(j)} = 0$ or $y^{(j)} = 1$. The payoff of player j is

$$\pi^{(j)} = y^{(j)}\big(x^{(j)\prime}\delta^{(j)} + \beta^{(j)}y^{(-j)} + u^{(j)}\big), \tag{3}$$

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 $\theta = (\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)})$:

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in S_{\{(0,0)\}|x;\theta} \equiv \{u : u^{(j)} < -x^{(j)\prime}\delta^{(j)}, j = 1,2\}, \\ \{(0,1)\} & u \in S_{\{(0,1)\}|x;\theta} \equiv \{u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}\}, \\ & \quad \cup\{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} > x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\} \\ \{(1,0)\} & u \in S_{\{(1,0)\}|x;\theta} \equiv \{u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ & \quad \cup\{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \\ \{(1,1)\} & u \in S_{\{(1,1)\}|x;\theta} \equiv \{u : u^{(j)} > -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, j = 1,2\}, \\ \{(1,0),(0,1)\} & u \in S_{\{(0,1),(1,0)\}|x;\theta} \equiv \{u : -x^{(j)\prime}\delta^{(j)} < u^{(j)} < -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, j = 1,2\}. \end{cases} \tag{4}$$

Note that the model predicts multiple equilibria when $u \in S_{\{(0,1),(1,0)\}|x;\theta}$ (see Figure 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

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in S_{\{(0,0)\}|x;\theta} \equiv \{u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \\ & \quad \cup\{-x^{(1)\prime}\delta^{(1)} - \beta^{(1)} \le u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ \{(0,1)\} & u \in S_{\{(0,1)\}|x;\theta} \equiv \{u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \ge -x^{(2)\prime}\delta^{(2)}\}, \\ \{(1,0)\} & u \in S_{\{(1,0)\}|x;\theta} \equiv \{u^{(1)} \ge -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}, \\ \{(1,1)\} & u \in S_{\{(1,1)\}|x;\theta} \equiv \{u^{(1)} \ge -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \ge -x^{(2)\prime}\delta^{(2)}\} \\ & \quad \cup\{u^{(1)} \ge -x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \le u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}, \\ \{(0,0),(1,1)\} & u \in S_{\{(0,0),(1,1)\}|x;\theta} \equiv \{-x^{(j)\prime}\delta^{(j)} - \beta^{(j)} \le u^{(j)} < -x^{(j)\prime}\delta^{(j)},\ j = 1,2\}. \end{cases} \tag{5}$$

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

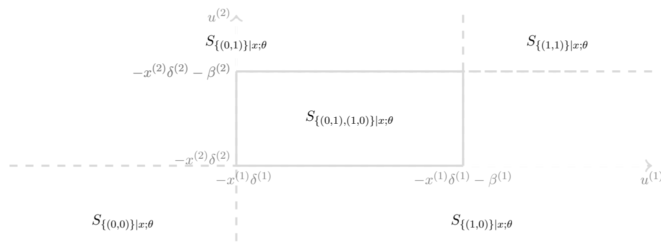

Figure 1: Level sets of $G(\cdot|x;\theta)$ with $\beta^{(j)} < 0, j = 1, 2$.

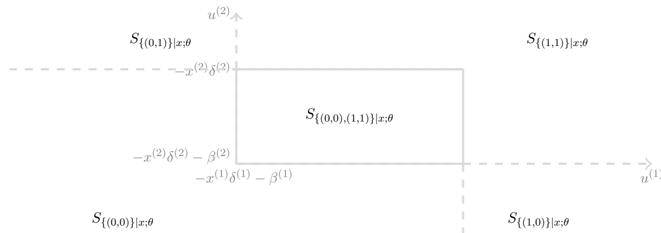

Figure 2: Level sets of $G(\cdot|x;\theta)$ with $\beta^{(j)} > 0, j = 1, 2$.

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

Example 2 (Heterogeneous Choice Sets): Consider a discrete choice model, with a finite universe of alternatives $\mathcal{J} = \{1, \dots, J\}$. 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:

$$Y \in \arg\max_{j \in C} W(j, X, U; \theta). \tag{6}$$

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 (6) then induces a unique likelihood function.[2] Alternatively, Barseghyan et al. (2021) only assume each decision 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

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

is a measurable correspondence:

$$G(U|x;\theta) = \cup_{K\subseteq\mathcal{J}:|K|\geq\kappa}\{\arg\max_{j\in K} W(j,x,U;\theta)\}. \tag{7}$$

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

$$\nu_\theta(A|x) = \int_{\mathcal{U}} 1\{G(u|x;\theta) \subseteq A\}dF_\theta(u), \quad \forall A \in \mathcal{C}. \tag{8}$$

This functional uniquely determines the distribution of the random set $G|X$ (Molchanov, 2005). The model prediction (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 (1) is the *core* of ν_θ:

$$\text{core}(\nu_\theta(\cdot|x)) \equiv \{Q \in \mathcal{M}(\Sigma_Y, \mathcal{X}) : Q(A|x) \geq \nu_\theta(A|x), A \in \mathcal{C}\}. \tag{9}$$

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 (9), one can define the set of conditional densities associated with $\text{core}(\nu_\theta(\cdot|x))$:

$$\mathfrak{q}_\theta \equiv \left\{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu,\ Q \in \text{core}(\nu_\theta(\cdot|x)),\ x \in \mathcal{X}\right\}. \tag{10}$$

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

$$I(f||f') \equiv \int_S f \ln \frac{f}{f'} d\zeta. \tag{11}$$

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

$$I(f||\mathfrak{f}) \equiv \inf_{f' \in \mathfrak{f}} I(f||f') \tag{12}$$

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

$$I(f||f') \equiv \int_{\mathcal{Y}\times\mathcal{X}} f(y, x) \ln \frac{f(y|x)}{f'(y|x)} d\zeta(y, x) \tag{13}$$

and use Eq. (13) in the KL divergence measure in Eq. (12).

We aim to test the following null hypothesis:

$$H_0 : I(p_0||\mathfrak{Q}_1) = I(p_0||\mathfrak{Q}_2). \tag{14}$$

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

$$H_1 : I(p_0||\mathfrak{Q}_1) < I(p_0||\mathfrak{Q}_2). \tag{15}$$

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 (1996). 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:

$$\begin{aligned} I(p_0||\mathfrak{q}_\theta) &= \inf_{q \in \mathfrak{q}_\theta} \int_{\mathcal{Y}\times\mathcal{X}} p_0(y, x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\zeta(y, x) \\ &= \int_{\mathcal{X}} p_{0,x}(x) \inf_{q_{y|x} \in \mathfrak{q}_{\theta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y) d\nu(x), \end{aligned} \tag{16}$$

where

$$\mathfrak{q}_{\theta,x} = \{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu,\ Q \in \text{core}(\nu_\theta(\cdot|x))\}. \tag{17}$$

The inner optimization problem in (16) is a convex program with a strictly convex objective function. Hence, a unique solution exists. Let

$$q^*_{\vartheta,y|x}(\cdot|x;p_{0,y|x}) = \underset{q_{y|x}\in\mathfrak{q}_{\vartheta,x}}{\arg\min}\int_{\mathcal{Y}} p_{0,y|x}(y|x)\ln\frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)}d\mu(y). \tag{18}$$

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 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 (2022).

Definition 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.*

Hence, a model is misspecified when one cannot recover the data-generating process p_0 even if one augments the model by a selection mechanism. For conceptual purposes, it is

[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.

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

Definition 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* $(y, x) \in \mathcal{Y} \times \mathcal{X}$.

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 $\{(Y_i, X_i), i = 1, \dots, n\}$ 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. 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 $i \in \{1, \dots, n\}$) into two equal halves denoted by I_1 and I_2. For each $\ell \in \{1, 2\}$, let $I_{-\ell} = \{1, \dots, n\} \setminus I_\ell$.

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

$$\hat{T}_{I_\ell} = \frac{\sqrt{n/2}\,\widehat{\mathrm{QLR}}_{I_\ell} + \hat{\omega}_{I_\ell} U_\ell}{\sqrt{\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) + \hat{\omega}^2_{I_\ell}}}, \tag{19}$$

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

$$\hat{\omega}_{I_\ell} = (1 + C \cdot \ln n \cdot \hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}))^{-1} \tag{20}$$

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

$$H_1 : E_{P_0}[L(X, \theta_1^*, p_{0,y|x})] > E_{P_0}[L(X, \theta_2^*, p_{0,y|x})]. \tag{21}$$

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 $\widehat{\text{QLR}}_{I_\ell}$, 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 $\widehat{\text{QLR}}_{I_\ell}$'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 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 data-generating process p_0. It reduces to a singleton containing the *pseudo-true parameter value* θ_s^* as in White (1996) 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

$$\theta_s^*(\theta_s, P_0) = \underset{\theta_s^* \in \Theta_s^*(P_0)}{\arg\inf} \|\theta_s - \theta_s^*\|. \tag{22}$$

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 $\mathcal{F}^k = \{f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | f(y, x) = \prod_{s=1}^2 (\ln q^*_{\theta_s, y|x}(y|x; p_{y|x}))^{k_s},\ \theta \in \Theta,\ p_{y|x} \in \mathcal{H}\}$.

3.1 The asymptotic size and power of the proposed test

Throughout, we assume the availability of a random sample.

Assumption 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 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 $\mathcal{A}_G = \operatorname{supp}(G(\cdot|x; \theta_s)) \equiv \{A \subseteq \mathcal{Y} : F_{\theta_s}(G(U|x; \theta_s) = A) > 0\}$ 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 2(a) restricts attention to models with discrete outcomes. Assumption 2(b) requires the support of the correspondence $G(\cdot|x; \theta_s)$ not to vary with $\theta_s \in \Theta_s$. Assumption 2(c) is easily verified when F_{θ_s} is differentiable in θ_s. Under Assumption 2, we can establish the differentiability of $L(x, \theta_s, p_{y|x})$ with respect to θ_s (see Lemma 1(i) in Appendix). 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: *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 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(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(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(c) imposes Lipschitz continuity on $E_{P_0}[m(X,\theta_s,p_{y|x})]$ with respect to $p_{y|x}$. Assumption 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 1–4, 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 2 in Appendix).

Next, we analyze the asymptotic behavior of the subsample QLR statistic $\widehat{\mathrm{QLR}}_{I_\ell}$ for $\ell \in \{1,2\}$. Under Assumption 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 1(ii) in Appendix). We add the following regularity conditions.

Assumption 5: *(a) There exists a function* $B(\cdot):\mathcal{X}\to\mathbb{R}$ *such that* $E[B^2(X)]<\infty$ *and for each* $s \in \{1,2\}$, $\sup_{\theta_s\in\Theta_s,p_{y|x}\in\mathcal{H}}\|m(x,\theta_s,p_{y|x})\| \leq B(x)$, $\sup_{\theta_s\in\Theta_s,p_{y|x},\tilde{p}_{y|x}\in\mathcal{H}}|D(x,\theta_s,p_{y|x},\tilde{p}_{y|x}-p_{y|x})| \leq B(x)$, *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,*

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

(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 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 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 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 1–5, we can obtain an asymptotically linear representation of each model's contribution to the subsample QLR statistic (see Lemma 3 in Appendix):

$$(n/2)^{-1/2} \sum_{i \in I_\ell} L(X_i, \hat{\theta}_{s, I_{-\ell}}, \hat{p}_{I_\ell, y|x}) = (n/2)^{-1/2} \sum_{i \in I_\ell} \ln q^*_{\theta_s^*(\hat{\theta}_{s, I_{-\ell}}, P_0), y|x}(Y_i|X_i; p_{0,y|x}) + o_p(1). \tag{23}$$

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 (23) to the two models, we may approximate the subsample QLR statistic as follows[5]

$$\begin{aligned} &\sqrt{n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_0}) \\ &\quad = \sqrt{2/n} \sum_{i \in I_\ell} (\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_0)}(Y_i|X_i; p_{0,y|x}) - E_{P_0}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_0)}(Y|X; p_{0,y|x})]) + o_p(1). \end{aligned} \tag{24}$$

To ensure the asymptotic normality of the leading term in (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 4 in Appendix).

Assumption 6: *There exist positive constants* M *and* ϵ *and a function* $D : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ *such that* $E_{P_0}[|D(Y, X)|^{2+\epsilon}] \leq M$ *and for all* $(y, x) \in \mathcal{Y} \times \mathcal{X}$ *and* $\theta^* \in \Theta^*(P_0)$, $|\lambda_{\theta^*}(y|x; p_{0,y|x}) - E_{P_0}[\lambda_{\theta^*}(Y|X; p_{0,y|x})]| \leq D(y, x)\sigma_{P_0}(\theta^*)$.

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

Let $\sigma^2_{P_0}(\theta) = E_{P_0}[\lambda^2_\theta(Y|X; p_{0,y|x})] - E_{P_0}[\lambda_\theta(Y|X; p_{0,y|x})]^2$. The asymptotic variance of the leading term in (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 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 $k_1 + k_2 \leq 2$, $\sup_{f\in\mathcal{F}^k} |\mathbb{G}_n(f)| = O_p(1)$, 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 $\theta, \tilde\theta \in \Theta$, $|E_{P_0}[\lambda^k_\theta(Y|X; p_{y|x})] - E_{P_0}[\lambda^k_{\tilde\theta}(Y|X; \tilde p_{y|x})]| \leq K_\lambda(\|\theta - \tilde\theta\| + \|p_{y|x} - \tilde p_{y|x}\|_{\mathcal{H}})$, $k = 1, 2$.*

Assumption 7(a) bounds the first moment of the profiled log-likelihood. Assumption 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 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 1–7, 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 5 in Appendix).

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 (24), causing the asymptotic distribution of $\sqrt{n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_0})$ 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 5: *Let $\mathcal{P}$ be the set of DGPs such that Assumptions 1–7 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 (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 (20) has the following property (see Lemmma 6 in Appendix):[6]

CONDITION 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 $H_0 : \mathrm{QLR}_{P_0} = 0$. Define the two-sided model-selection test of

[6] Any other sequences that satisfy Condition 1 can be used.

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.

THEOREM 1: *Suppose Assumptions 1–7 hold. Then, for any sequence $\{P_n \in \mathcal{P}_0\}$,*

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

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

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

THEOREM 2: *Suppose Assumptions 1–7 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 $\sqrt{n}\,\text{QLR}_{P_n} \to h \in (0,\infty)$,*

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

4 Examples

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

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 (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 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]:

$$q^*_{\theta,y|x}((0,0)|x;p_{0,y|x}) = F_\theta(S_{\{(0,0)\}|x;\theta}|x) \tag{25}$$
$$q^*_{\theta,y|x}((1,1)|x;p_{0,y|x}) = F_\theta(S_{\{(1,1)\}|x;\theta}|x) \tag{26}$$
$$q^*_{\theta,y|x}((1,0)|x;p_{0,y|x}) = p_{0,M}((1,0)|x)\eta^1(\theta;x)\mathbb{I}^1(x;\theta) + \eta^2(\theta;x)\mathbb{I}^2(x;\theta) + \eta^3(\theta;x)\mathbb{I}^3(x;\theta), \tag{27}$$

where

$$\mathbb{I}^1(x;\theta) \equiv 1\big\{\eta_1^3(\theta;x)/\eta^1(\theta;x) \le p_{0,M}((1,0)|x) \le \eta^2(\theta;x)/\eta^1(\theta;x)\big\}$$
$$\mathbb{I}^2(x;\theta) \equiv 1\big\{p_{0,M}((1,0)|x) > \eta^2(\theta;x)/\eta^1(\theta;x)\big\}$$
$$\mathbb{I}^3(x;\theta) \equiv 1\big\{p_{0,M}((1,0)|x) < \eta^3(\theta;x)/\eta^1(\theta;x)\big\}.$$

Let us explain the intuition behind (25)-(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 (27) is given by the upper bound $\eta^2(\theta;x)$ of the predicted probability 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 (27) is given by the lower bound $\eta^3(\theta;x)$ of the probability of $Y=(1,0)$.

[7]See Kaido and Molinari (2022) for derivation. Since $\mathcal{Y}$ consists of four outcomes, we report the value of the profiled likelihood for three outcome values below.

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 $N \in \{0, 1, 2\}$:

$$q^*_{\theta,y|x}(0|x; p_{0,y|x}) = F_\theta(S_{\{(0,0)\}|x;\theta}|x), \tag{28}$$

$$q^*_{\theta,y|x}(1|x; p_{0,y|x}) = F_\theta(S_{\{(0,1)\}|x;\theta}|x) + F_\theta(S_{\{(0,1),(1,0)\}|x;\theta}|x) + F_\theta(S_{\{(1,0)\}|x;\theta}|x), \tag{29}$$

$$q^*_{\theta,y|x}(2|x; p_{0,y|x}) = F_\theta(S_{\{(1,1)\}|x;\theta}|x), \tag{30}$$

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 (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 outcomes $(1, 1)$ out of the "Non-Monopoly" event $Y \in \{(0, 0), (1, 1)\}$. An argument similar to structure 1 shows the profiled-likelihood is

$$q^*_{\theta,y|x}((1,0)|x; p_{0,y|x}) = F_\theta(S_{\{(1,0)\}|x;\theta}|x) \tag{31}$$

$$q^*_{\theta,y|x}((0,1)|x; p_{0,y|x}) = F_\theta(S_{\{(0,1)\}|x;\theta}|x) \tag{32}$$

$$\begin{aligned} q^*_{\theta,y|x}((0,0)|x; p_{0,y|x}) &= p_{0,N}((0,0)|x)\eta^1(\theta_2; x)\mathbb{I}^1(x;\theta) \\ &\quad + [\eta^1(\theta;x) - \eta^2(\theta;x)]\mathbb{I}^2(x;\theta) + [\eta^1(\theta;x) - \eta^3(\theta;x)]\mathbb{I}^3(x;\theta), \end{aligned} \tag{33}$$

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

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

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

$$W(L_j; U) = \mu v(-\pi_j - c_j; U) + (1-\mu)v(-\pi_j; U). \tag{34}$$

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

$$F_{C|X,\theta}(K|x) \equiv P(C = K|X = x) = \prod_{l \in K} \phi_l(x) \prod_{k \notin K} (1 - \phi_k(x)), \tag{35}$$

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

$$f_\theta(j|x, K) = \int 1\{W(L_j; u) > W(L_k; u), \forall k \in K, k \neq j\} dF_{U|X,\theta}(u) \tag{36}$$

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:

$$q^*_{\theta_1,y|x}(j|x) = \sum_{K \in \mathcal{C}_j} \prod_{l \in K} \phi_l(x) \prod_{k \notin K} (1 - \phi_k(x)) f_\theta(j|x, K), \ j = 1, \ldots, J. \tag{37}$$

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 (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

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

utility functions with negligible third derivative (NTD), i.e.,

$$\frac{v(w+\Delta)}{v'(w)} - \frac{v(w)}{v'(w)} = \Delta - \frac{U}{2}\Delta^2. \tag{38}$$

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) 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

$$G(U|X;\theta) = \begin{cases} \{2,3\} & \text{if } U < \bar{\tau}(X) \\ \{1,2\} & \text{if } U \geq \bar{\tau}(X). \end{cases} \tag{39}$$

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

$$Q(\{2,3\}|x) \geq \nu_\theta(\{2,3\}|x) = F_\theta(G(U|X;\theta) \subseteq \{2,3\}|x) = F_\theta(U < \bar{\tau}(x)) \tag{40}$$

$$Q(\{1,2\}|x) \geq \nu_\theta(\{1,2\}|x) = F_\theta(G(U|X;\theta) \subseteq \{1,2\}|x) = F_\theta(U \geq \bar{\tau}(x)). \tag{41}$$

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 (16) yields the following profiled-likelihood:

$$q^*_{\theta_2,y|x}(1|x) = p_{0,y|x}(1|x)\mathbb{I}^1(x;\theta) + p_{0,1|12}(1|x)\eta(\theta;x)\mathbb{I}^2(x;\theta) + \eta(\theta;x)\mathbb{I}^3(x;\theta) \tag{42}$$

$$q^*_{\theta_2,y|x}(3|x) = p_{0,y|x}(3|x)\mathbb{I}^1(x;\theta) + [1-\eta(\theta;x)]\mathbb{I}^2(x;\theta) + p_{0,3|23}(1|x)[1-\eta(\theta;x)]\mathbb{I}^3(x;\theta), \tag{43}$$

where

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

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

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 $X_2^{(j)}$ 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 data generating processes:

- DGP1: $X_2^{(j)} \overset{i.i.d.}{\sim} \text{Bernoulli}(0.5)$.
- 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:

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

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

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

where $\widehat{\text{LR}}_n$ 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 1–3 and Figure 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 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 $\widehat{\text{LR}}_n$, 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.

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 1: Rejection Probabilities ($n = 1000$)

[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.

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
$\hat{T}_n^{\mathrm{HS}}$	0.057	0.056	0.058	0.061	0.066	0.088	0.147	0.262	0.430	0.609	0.751

Table 2: Rejection Probabilities ($n = 500$)

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
$\hat{T}_n^{\mathrm{HS}}$	0.090	0.090	0.092	0.095	0.120	0.177	0.289	0.446	0.610	0.765	0.882

Table 3: Rejection Probabilities ($n = 250$)

6 Concluding remarks

This paper 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.

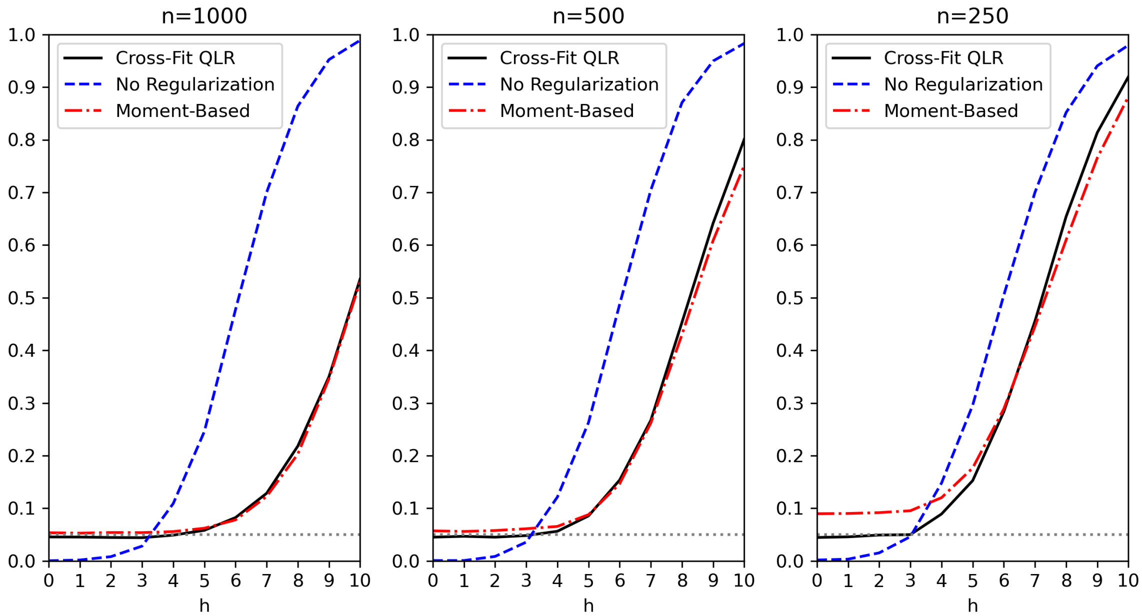

Figure 3: Power Curves

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

Table 4: Average Runtime (in sec.)

A Proofs of Theorems 1 and 2

Proof of Theorem 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 3 and 4,

$$
\begin{aligned}
&\sqrt{a_n/2}(\widehat{\mathrm{QLR}}_{I_\ell} - \mathrm{QLR}_{P_{a_n}}) \\
&= o_p(1) + (a_n/2)^{-1/2} \sum_{i\in I_\ell} (\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})}(Y_i|X_i; p_{a_n,y|x}) - E_{P_{a_n}}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})}(Y|X; p_{a_n,y|x})]) \\
&= o_p(1) + \sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot (Z_\ell + o_p(1)) \\
&= \sigma_{P_{a_n}}(\theta^*(\hat{\theta}_{I_{-\ell}}, P_{a_n})) \cdot Z_\ell + o_p(1).
\end{aligned} \tag{44}
$$

We consider two cases.

Case 1: $\sigma_\infty = 0$. By Lemma 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 (44), $\sqrt{a_n/2}\widehat{\mathrm{QLR}}_{I_\ell} = o_p(1) \cdot O_p(1) + o_p(1) = o_p(1)$. By Condition 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 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 (44) and Condition 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 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 1 and consider two cases.

Case 1: $\sigma_\infty = 0$. By Lemma 5, $\hat{\sigma}^2_{I_\ell}(\hat{\theta}_{I_{-\ell}}) = o_p(1)$. Then, by (44) and Condition 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 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 (44) and Condition 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,

$$\begin{aligned}
\lim_{n\to\infty} E_{P_{a_n}}[\phi_{a_n}^{\text{2-sided}}(\alpha)] &= \lim_{n\to\infty} \Pr_{P_{a_n}}(|\hat{T}_{a_n}| > z_{1-\alpha/2}) \\
&\geq \Pr(|Z + h/(\omega_\infty \vee \sigma_\infty)| > z_{1-\alpha/2}) \\
&= 1 - \Phi(z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)) + \Phi(-z_{1-\alpha/2} - h/(\omega_\infty \vee \sigma_\infty)), \\
\lim_{n\to\infty} E_{P_{a_n}}[\phi_{a_n}^{\text{1-sided}}(\alpha)] &= \lim_{n\to\infty} \Pr_{P_{a_n}}(\hat{T}_{a_n} > z_{1-\alpha}) \\
&\geq \Pr(Z + h/(\omega_\infty \vee \sigma_\infty) > z_{1-\alpha}) \\
&= 1 - \Phi(z_{1-\alpha} - h/(\omega_\infty \vee \sigma_\infty)).
\end{aligned}$$

Therefore, the desired result follows. □

B Auxiliary Lemmas and their proofs

LEMMA 1: *Suppose that Assumption 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 1. For part (i), note that by definition,

$$\begin{aligned}
L(x,\theta_s,p_{y|x}) = \max_{q\in\Delta} &\sum_{y\in\mathcal{Y}} p_{y|x}(y|x) \ln q(y) \\
&s.t. \quad \nu_{\theta_s}(A|x) \leq \sum_{y\in A} q(y), \; A \in \mathcal{C}.
\end{aligned} \tag{45}$$

Hence, the desired result follows from Theorem 3.1(i) of Kaido and Molinari (2022) 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 (45) has a unique solution $q^*_{\theta_s,y|x}(\cdot|x;p_{y|x})$. □

LEMMA 2: *Suppose that Assumptions 1–4 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 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

$$n^{d_p} Q_n(\theta) = n^{d_p} E_{P_0}[m(X, \theta, p_{0,y|x})] + n^{d_p - 1/2} \mathbb{G}_n(m(\cdot, \theta, \hat{p}_{n,y|x})) \\ + n^{d_p}(E_{P_0}[m(X, \theta, \hat{p}_{n,y|x})] - E_{P_0}[m(X, \theta, p_{0,y|x})]).$$

By Assumption 3(b), $\mathbb{G}_n(m(\cdot, \theta, \hat{p}_{n,y|x})) = O_p(1)$. By Assumptions 3(c) and 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(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 3: *Suppose Assumptions 1–5 hold. Then for each $s, \ell \in \{1, 2\}$,*

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

Proof of Lemma 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 5(a), for any $\theta,\tilde{\theta}\in\Theta$ and $p_{y|x},\tilde{p}_{y|x}\in\mathcal{H}$,

$$
|L(x,\theta,p_{y|x}) - L(x,\tilde{\theta},\tilde{p}_{y|x})| \le B(x)(\|\theta-\tilde{\theta}\| + \|p_{y|x}-\tilde{p}_{y|x}\|_{\mathcal{H}}).
$$

Hence, by Assumption 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 $\delta_n = o(1)$,

$$
\sup_{\|\tilde{\theta}_s-\theta_s\|+\|\tilde{p}_{y|x}-p_{y|x}\|_{\mathcal{H}}\le\delta_n} |\mathbb{G}_n(L(\cdot,\tilde{\theta}_s,\tilde{p}_{y|x}) - L(\cdot,\theta_s,p_{y|x}))| = o_p(1).
$$

By Lemma 2, $\|\hat{\theta}_n - \theta^*(\hat{\theta}_n,P_0)\| = O_p(n^{-d_p})$. Hence, for all $\delta_n = o(1)$, $\|\hat{\theta}_n - \theta^*(\hat{\theta}_n,P_0)\| \le \delta_n$ with probability approaching 1. Similarly, by Assumption 4, $\|\hat{p}_{n,y|x} - p_{0,y|x}\|_{\mathcal{H}} \le \delta_n$ 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

$$
\begin{aligned}
&\sqrt{n}(E_{P_0}[L(X,\hat{\theta}_n,\hat{p}_{n,y|x})] - E_{P_0}[L(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})]) \\
&= \sqrt{n}(\hat{\theta}_n - \theta^*(\hat{\theta}_n,P_0))' E_{P_0}[m(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x})] \\
&\quad + \sqrt{n}E_{P_0}[D(X,\theta^*(\hat{\theta}_n,P_0),p_{0,y|x},\hat{p}_{n,y|x}-p_{0,y|x})] + \sqrt{n}E_{P_0}[r(X,\hat{\theta}_n,\hat{p}_{n,y|x})],
\end{aligned}
$$

where $r(x,\theta,p_{y|x}) \equiv L(x,\theta,p_{y|x}) - L(x,\theta^*(\theta,P_0),p_{0,y|x}) - (\theta-\theta^*(\theta,P_0))'m(x,\theta^*(\theta,P_0),p_{0,y|x}) - D(x,\theta^*(\theta,P_0),p_{0,y|x},p_{y|x}-p_{0,y|x})$. 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 2, and Assumptions 4 and 5(a),

$$\begin{aligned}\sqrt{n}|E_{P_0}[r(X,\hat{\theta}_n,\hat{p}_{n,y|x})]| &\le \sqrt{n}E_{P_0}[|r(X,\hat{\theta}_n,\hat{p}_{n,y|x})|]\\ &\le E[B(X)]\sqrt{n}(\|\hat{\theta}_n-\theta^*(\hat{\theta}_n,P_0)\|^2+\|\hat{p}_{n,y|x}-p_{0,y|x}\|_{\mathcal{H}}^2)=o_p(1).\end{aligned}$$

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 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 5(c). □

Lemma 4: *Suppose Assumptions 1–6 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 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 $n\in\mathbb{N}_+$,

$$\begin{aligned}\sum_{i\in I_\ell}E_{P_n}[|Z_{ni}|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}]&=\frac{\sum_{i\in I_\ell}E_{P_n}[|\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}},P_n)}(Y_i|X_i;p_{n,y|x})-E_{P_n}[\lambda_{\theta^*(\hat{\theta}_{I_{-\ell}},P_n)}(Y|X;p_{n,y|x})]|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}]}{(\sqrt{n/2}\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}},P_n)))^{2+\epsilon}}\\&\le\frac{(n/2)E_{P_n}[|D(Y,X)\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}},P_n))|^{2+\epsilon}]}{(\sqrt{n/2}\sigma_{P_n}(\theta^*(\hat{\theta}_{I_{-\ell}},P_n)))^{2+\epsilon}}\\&=(n/2)^{-\epsilon/2}E_{P_n}[|D(Y,X)|^{2+\epsilon}],\end{aligned}$$

where the inequality follows from the independence between I_1 and I_2 and Assumption 6. Hence, there exist $M,\epsilon>0$ such that for each $n\in\mathbb{N}_+$, $\sum_{i\in I_\ell}E_{P_n}[|Z_{ni}|^{2+\epsilon}|\hat{\theta}_{I_{-\ell}}]\le(n/2)^{-\epsilon/2}M$. By the law of iterated expectations, the Lyapounov condition holds:

$$\sum_{i\in I_\ell}E_{P_n}[|Z_{ni}|^{2+\epsilon}]\le(n/2)^{-\epsilon/2}M\to 0.$$

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. (2013), 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 5: *Suppose Assumptions 1–7 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 5. We omit the sample-splitting feature for readability. Define

$$V(\theta, p_{y|x}) = E_{P_0}[\lambda_\theta^2(Y|X; p_{y|x})] - E_{P_0}[\lambda_\theta(Y|X; p_{y|x})]^2,$$
$$V_n(\theta, p_{y|x}) = \frac{1}{n}\sum_{i=1}^n \lambda_\theta(Y_i|X_i; p_{y|x}) - \Big(\frac{1}{n}\sum_{i=1}^n \lambda_\theta(Y_i|X_i; p_{y|x})\Big)^2.$$

We can write

$$(\hat\sigma_n(\hat\theta_n))^2 - \sigma_{P_0}^2(\theta^*(\hat\theta_n, P_0)) = (V_n(\hat\theta_n, \hat p_{n,y|x}) - V(\hat\theta_n, \hat p_{n,y|x})) + (V(\hat\theta_n, \hat p_{n,y|x}) - V(\theta^*(\hat\theta_n, P_0), p_{0,y|x})).$$

We examine each term on the right-hand side. First, by Assumption 7(b),

$$\begin{aligned}
V_n(\hat\theta_n, \hat p_{n,y|x}) - V(\hat\theta_n, \hat p_{n,y|x}) &= n^{-1/2}\mathbb{G}_n(\lambda^2_{\hat\theta_n}(\cdot|\cdot; \hat p_{n,y|x})) - n^{-1/2}\mathbb{G}_n(\lambda_{\hat\theta_n}(\cdot|\cdot; \hat p_{n,y|x}))\\
&\quad \times (2E_{P_0}[\lambda_{\hat\theta_n}(Y|X; \hat p_{n,y|x})] + n^{-1/2}\mathbb{G}_n(\lambda_{\hat\theta_n}(\cdot|\cdot; \hat p_{n,y|x})))\\
&= O_p(n^{-1/2}) - O_p(n^{-1/2})\cdot(O_p(1) + O_p(n^{-1/2}))\\
&= O_p(n^{-1/2}).
\end{aligned}$$

Second, by the triangle inequality and Jensen's inequality,

$$\begin{aligned}
|V(\hat\theta_n, \hat p_{n,y|x}) - V(\theta^*(\hat\theta_n, P_0), p_{0,y|x})| &\le |E_{P_0}[\lambda^2_{\hat\theta_n}(Y|X; \hat p_{n,y|x})] - E_{P_0}[\lambda^2_{\theta^*(\hat\theta_n, P_0)}(Y|X; p_{0,y|x})]|\\
&\quad + 2|E_{P_0}[\lambda_{\hat\theta_n}(Y|X; \hat p_{n,y|x})] - E_{P_0}[\lambda_{\theta^*(\hat\theta_n, P_0)}(Y|X; p_{0,y|x})]|\\
&\quad \times \Big(\sum_{s=1}^2 \sup_{\theta_s\in\Theta_s, p_{y|x}\in\mathcal{H}} E_{P_0}[|\ln q^*_{\theta_s, y|x}(Y|X; p_{y|x})|]\Big).
\end{aligned}$$

By Lemma 2 and Assumptions 4 and 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 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

$$|V(\hat\theta_n, \hat p_{n,y|x}) - V(\theta^*(\hat\theta_n, P_0), p_{0,y|x})| \le O_p(n^{-d_p}) + O_p(n^{-d_p})\cdot O_p(1) = O_p(n^{-d_p}).$$

Putting everything together yields the desired result. □

Lemma 6: *Suppose that Assumptions 1–7 hold. Then, for each $\ell \in \{1, 2\}$, $\hat\omega_{I_\ell}$ defined in (20) satisfies Condition 1.*

Proof of Lemma 6. Fix $\ell \in \{1, 2\}$. To check Condition 1(a), note that by Lemma 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 1(b), note that by Lemma 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," .

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2021): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048.

Barseghyan, L., F. Molinari, T. O'Donoghue, and J. C. Teitelbaum (2013): "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103, 2499–2529.

Barseghyan, L., J. Prince, and J. C. Teitelbaum (2011): "Are Risk Preferences Stable across Contexts? Evidence from Insurance Data," *American Economic Review*, 101, 591–631.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp Identification Regions in Models With Convex Moment Predictions," *Econometrica*, 79, 1785–1821.

Berry, S. T. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 60, 889–917.

Berry, S. T. and G. Compiani (2022): "An Instrumental Variable Approach to Dynamic Models," *The Review of Economic Studies*, 90, 1724–1758.

Bresnahan, T. F. and P. C. Reiss (1990): "Entry in monopoly market," *The Review of Economic Studies*, 57, 531–553.

Caballero, R. J. and E. M. R. A. Engel (1999): "Explaining Investment Dynamics in U.S. Manufacturing: A Generalized (S, s) Approach," *Econometrica*, 67, 783–826.

Cameron, S. V. and J. J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, 262–333.

Chen, S. and H. Kaido (2023): "Robust Tests of Model Incompleteness in the Presence of Nuisance Parameters," .

Chen, X., H. Hong, and M. Shum (2007): "Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models," *Journal of Econometrics*, 141, 109–140, semiparametric methods in econometrics.

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

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

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81, 667–737.

Chesher, A. and A. M. Rosen (2017): "Generalized instrumental variable models," *Econometrica*, 85, 959–989.

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

Ciliberto, F. and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828.

Coate, S. and M. Conlin (2004): "A Group Rule-Utilitarian Approach to Voter Turnout: Theory and Evidence," *American Economic Review*, 94, 1476–1504.

Cohen, A. and L. Einav (2007): "Estimating Risk Preferences from Deductible Choice," *American Economic Review*, 97, 745–788.

Dickstein, M. J. and E. Morales (2018): "What do Exporters Know?*," *The Quarterly Journal of Economics*, 133, 1753–1801.

Eizenberg, A. (2014): "Upstream Innovation and Product Variety in the U.S. Home PC Market," *The Review of Economic Studies*, 81, 1003–1045.

Fack, G., J. Grenet, and Y. He (2019): "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions," *American Economic Review*, 109, 1486–1529.

Fafchamps, M. (1993): "Sequential Labor Decisions Under Uncertainty: An Estimable Household Model of West-African Farmers," *Econometrica*, 61, 1173–1197.

Fiacco, A. V. and Y. Ishizuka (1990): "Sensitivity and stability analysis for nonlinear programming," *Annals of Operations Research*, 27, 215–235.

Francois, P., I. Rainer, and F. Trebbi (2015): "How Is Power Shared in Africa?" *Econometrica*, 83, 465–503.

Galichon, A. and M. Henry (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78, 1264–1298.

Goeree, M. S. (2008): "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 76, 1017–1074.

Haile, P. A. and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111.

Honoré, B. E. and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

Hsu, Y.-C. and X. Shi (2017): "Model-selection tests for conditional moment restriction models," *The Econometrics Journal*, 20, 52–85.

Kaido, H. and F. Molinari (2022): "Information-Based Inference in Models with Set-Valued Predictions," Working Paper, Boston University and Cornell University.

——— (2024): "Information Based Inference In Models With Set-Valued Predictions And Misspecification," ArXiv:2401.11046.

Kaido, H., F. Molinari, and J. Stoye (2022): "CONSTRAINT QUALIFICATIONS IN PARTIAL IDENTIFICATION," *Econometric Theory*, 38, 596–619.

Kendall, C., T. Nannicini, and F. Trebbi (2015): "How Do Voters Respond to Information? Evidence from a Randomized Campaign," *American Economic Review*, 105, 322–53.

Li, T. (2009): "Simulation based selection of competing structural econometric models," *Journal of Econometrics*, 148, 114–123.

Liao, Z. and X. Shi (2020): "A nondegenerate Vuong test and post selection confidence intervals for semi/nonparametric models," *Quantitative Economics*, 11, 983–1017.

Miyauchi, Y. (2016): "Structural estimation of pairwise stable networks with nonnegative externality," *Journal of Econometrics*, 195, 224 – 235.

Molchanov, I. S. (2005): *Theory of random sets*, vol. 19, Springer.

Molinari, F. (2020): "Microeconometrics with partial identification," *Handbook of econometrics*, 7, 355–486.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Nyarko, Y. and A. Schotter (2002): "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70, 971–1005.

Palfrey, T. R. and J. E. Prisbrey (1997): "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *The American Economic Review*, 87, 829–846.

Paulson, A. L., R. M. Townsend, and A. Karaivanov (2006): "Distinguishing Limited Liability from Moral Hazard in a Model of Entrepreneurship," *Journal of Political Economy*, 114, 100–144.

Rivers, D. and Q. Vuong (2002): "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 5, 1–39.

Schennach, S. M. and D. Wilhelm (2017): "A Simple Parametric Model Selection Test," *Journal of the American Statistical Association*, 112, 1663–1674.

Sheng, S. (2020): "A Structural Econometric Analysis of Network Formation Games Through Subnetworks," *Econometrica*, 88, 1829–1858.

Shi, X. (2015a): "Model selection tests for moment inequality models," *Journal of Econometrics*, 187, 1–17.

——— (2015b): "A nondegenerate Vuong test," *Quantitative Economics*, 6, 85–121.

Tamer, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *The Review of Economic Studies*, 70, 147–165.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," .

White, H. (1996): *Estimation, inference and specification analysis*, 22, Cambridge university press.