Robust Tests of Model Incompleteness in the Presence of Nuisance Parameters*

Shuowen Chen† Hiroaki Kaido‡

October 8, 2021

Abstract

In discrete choice models, whether the model makes a unique prediction or not is often tied to important features of the underlying model such as the interdependence of agents' preferences in models of social interaction or the endogeneity of treatment assignments in triangular systems of binary outcome and treatment variables. We provide a novel test of model incompleteness using a score-based statistic. Our test statistic remains computationally tractable even with a moderate number of nuisance parameters because they have to be estimated only in the restricted complete model. A Monte Carlo experiment shows the proposed test outperforms existing tests in terms of local power. An empirical application to a model of entry in the airline industry illustrates the computational feasibility of the method.

Key Words: Incomplete Models, Score tests, Subvector inference, Strategic interaction, Discrete Choice Models

*We thank Iván Fernández-Val, Jean–Jacques Forneron, Paul S. Koh, Pierre Perron, Marc Rysman, Zhongjun Qu, and Guang Zhang for helpful discussions and suggestions. Kaido gratefully acknowledges financial support through NSF Grant SES-2018498.

†Boston University. Email: swchen@bu.edu.

‡Boston University. Email: hkaido@bu.edu.

1 Introduction

Models of discrete choice are used widely. Commonly used models combine a theory of choice (e.g. utility maximization) that predicts a unique outcome value with distributional assumptions on latent variables to describe the conditional distribution of the outcome given observable covariates. When the researcher is willing to work only with weak assumptions or has limited knowledge of the data generating process, however, these models often end up predicting multiple outcome values, which we call an *incomplete prediction*. We consider a form of incompleteness summarized as follows; an observable discrete outcome variable Y is known to satisfy

$$Y \in G(u|X;\theta), \tag{1.1}$$

where G collects all outcome values that are compatible with the model given the unobserved and observed variables (u, X) and a structural parameter θ. This type of incompleteness arises in a variety of contexts. In single-agent discrete choice models, multiple outcomes are predicted when the agent's choice set is unobservable and consistent with a wide range of choice set formation processes (Barseghyan et al., 2021). In discrete games such as firms' market entry or household's labor supply decisions, multiple equilibria may exist but one may not know how an equilibrium outcome gets selected (Bresnahan and Reiss, 1991a; Ciliberto and Tamer, 2009). In panel dynamic discrete choice models, one's theory may be silent about how an initial observation is generated (Heckman, 1978; Honoré and Tamer, 2006). Recent empirical studies have fruitfully applied econometric methods for such incomplete models in different areas; they include English auctions (Haile and Tamer, 2003), strategic voting (Kawai and Watanabe, 2013), product offerings (Eizenberg, 2014; Wollmann, 2018), network formation (de Paula et al., 2018; Sheng, 2020), school choice (Fack et al., 2019), and major choice (Henry et al., 2020).

Whether a model needs to allow incompleteness to be consistent with data is a natural and important empirical question. Furthermore, whether a model is complete or not is often tied to key features of the underlying structural model. In a commonly used model of market entry, multiple equilibria exist only when the players interact strategically. Testing the presence of the strategic interaction effects and making inference on their signs can provide important information for policymaking (de Paula and Tang, 2012). In a binary choice model with a binary treatment assignment, taking a control function approach yields an incomplete model only if the treatment assignment is endogenous. When instrumental variables are available, one can test the potential endogeneity of the treatment assignment through the test of the model incompleteness.

This paper's goal is to develop a procedure for testing the model completeness against incompleteness. In leading examples, the null hypothesis of model completeness can be formulated as restrictions on a subvector of structural parameters. We, therefore, consider testing the presence of incompleteness by testing a hypothesis on a subvector. Inference on subvectors of parameters has attracted considerable attention in the recent econometrics literature on incomplete models (see Bugni et al., 2017; Kaido et al., 2019, and references therein). A key challenge surrounding subvector inference is the computational cost for implementing existing methods (Molinari, 2020, sec. 6). As we discuss below, progress has been made recently within a class of models having a particular structure. However, computational challenges remain for models outside such a class especially when there are multiple nuisance components, which is common in applications.

This paper develops a test that can be implemented easily even in the presence of nuisance parameters. We achieve this by focusing on a class of models that are *complete* when the subvector (called β) of interest is set to its null value. We then propose a novel score-based test statistic. Advantages of this approach are (i) the score statistic only requires estimation of nuisance parameters in the restricted model, which is complete; (ii) the nuisance parameters can be estimated by standard point estimators (e.g. restricted MLE) using package software; and (iii) the statistic's null distribution can be simulated easily.

The basic idea behind our tests is as follows. The model incompleteness, in general, implies multiple (typically infinitely many) likelihood functions, which makes challenging to apply standard likelihood-based tests. However, the class of models we consider has properties that make score-based inference tractable. First, under any value of structural parameter $\theta_0 + h$ violating the null hypothesis locally, there is a "least favorable" data generating density q_{θ_0+h} that is most difficult to distinguish from the density q_{θ_0} under the null hypothesis. We may then view the mapping $\theta \mapsto q_\theta$ as a "least favorable parametric model", along which detecting the deviation from θ_0 is most difficult. Second, one may consider a test that maximizes a measure of local discrimination (between θ_0 and $\theta_0 + h$) based on the least favorable parametric model. This leads to a robust test based on our score function. The test is designed to detect any local deviation from the null hypothesis no matter what the unknown selection mechanism is.

The score function typically depends on the nuisance components δ of the parameter vector. Exploiting the property that the model is complete under the null hypothesis, we show that a $\sqrt{n}$-consistent point estimator of δ can be constructed and plugged into the score. This procedure avoids evaluating the test statistic over a grid of nuisance parameters and makes our procedure computationally tractable. When the restricted maximum likelihood estimator of δ is used, the score-based statistic has a limiting distribution that can be easily simulated. When other estimators are used, we construct our test statistic using an orthogonalized version of the score following the insights of Neyman's $C(\alpha)$-test (Neyman, 1959, 1979). This makes the distribution of the statistic insensitive to the effects of the estimated nuisance parameters.

In sum, we address key issues surrounding tests of incompleteness and subvector inference by combining (i) the score function associated with the least favorable parametric model and (ii) a point estimator of the nuisance components. To our knowledge, this type of test for incomplete models is new. The key structure we exploit is that the model is complete under the null hypothesis, which can be restrictive in some settings. However, there are more general tests that can be applied to settings outside the scope of our paper with a higher computational cost (Bugni et al., 2017; Kaido et al., 2019). Our test hence complements them in a particular class of testing problems.

While this paper focuses on testing the model completeness, estimating some or all components of θ may be the ultimate goal of the researcher in some applications. In such a case, our restricted maximum likelihood estimator provides a consistent estimator of δ if the null hypothesis is true, whereas robust inference methods for subcomponents of the structural parameter in the literature can be used in case the alternative hypothesis is true. Our test, therefore, can be viewed as a specification test, which naturally raises a question regarding its impact on any post-model selection inference. While we defer a formal analysis to another work, we propose a hybrid procedure that aims at controlling the potential distortion of the model selection step using a shrinkage method borrowing the insights from the moment selection literature (Andrews and Soares, 2010; Romano et al., 2014).

1.1 Relation to the literature

Our paper belongs to the literature on identification and inference in incomplete models. The seminal work of Tamer (2003) showed an incomplete model induces multiple distributions and implies partially identifying restrictions on parameters. Recent developments in the literature (Galichon and Henry, 2011; Beresteanu et al., 2011; Chesher and Rosen, 2017) provided tools to systematically derive so-called *sharp identifying restrictions* (SIRs), which convert all model information into a set of equality and inequality restrictions on the conditional moments of the observables. Inference methods based on the sample analogs of such moment restrictions are extensively studied (see Canay and Shaikh, 2017, and references therein). Instead of using the sample moment restrictions, we use the original sharp identifying restrictions to derive the least favorable parametric model. Our test statistics are then constructed using the score function associated with the least favorable parametric model. This is akin to deriving a score function using a parametric specification in a complete model.

Hypothesis testing in incomplete models is studied extensively. As discussed earlier, many of them are based on the sample analogs of conditional or unconditional moment restrictions. There are attempts to improve the computational tractability of the moment-based inference methods within a special class of models. They include Andrews et al. (2019); Cox and Shi (2020) who assume that moment inequality restrictions implied by the model are linear conditional on some observable variables. This paper focuses on another special class, in which the model becomes complete under the null hypothesis. There are recent developments on inference methods based on likelihood-ratios (Chen et al., 2018; Kaido and Zhang, 2019) as well. Our approach builds on Kaido and Zhang (2019) (KZ19, henceforth) who used the least favorable pairs (LFPs) of distributions studied in the robust statistics literature to conduct likelihood-ratio (LR) tests in incomplete models. We use score functions derived from the LFPs and construct a statistic tailored to our testing problem.

This paper's framework can be used to test the presence of strategic interaction effects and of multiple equilibria in complete information games. Related problems are studied in other classes of models. For incomplete information games, de Paula and Tang (2012) introduced a semiparametric inference procedure on the signs of strategic interaction effects. For finite-state Markov games, Otsu et al. (2016) provide procedures to test whether the conditional choice probabilities, state transition and other features of games are homogeneous across cross-sectional units. Rejection of their null hypothesis could occur when multiple equilibria are present. In the context of network formation with a large number of agents, Pelican and Graham (2021) develop a procedure to test whether agents' preferences over networks are interdependent. Using a Logit specification, they propose conditional tests and introduce an MCMC algorithm to implement their test.

Finally, our framework can be applied to triangular systems involving a binary outcome and a binary endogenous variable. We show that taking a control function approach in such a setting leads to a model with an incomplete prediction. Namely, the model involves a set-valued control function. Our framework can be used to conduct a test of the endogeneity of the treatment assignment with weak assumptions. To our knowledge, this test is new to the literature and provides an alternative to the existing proposal by Wooldridge (2014) who makes additional high-level assumptions.

2 Set-up

Let Y be a discrete outcome taking values in a finite set $\mathcal{Y}$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ be a vector of observable covariates and let $u \in U \subseteq \mathbb{R}^{d_U}$ be a vector of unobservable variables. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a finite dimensional parameter.

The prediction of a structural model is summarized by a weakly measurable set-valued map $G : U \times \mathcal{X} \times \Theta \rightsquigarrow \mathcal{Y}$. We assume that Y takes one of the values in $G(u|X;\theta)$ with probability 1. The map G describes how the observable and unobservable characteristics of individuals and/or economic environments translate into a set of possible outcome values. It reflects restrictions imposed by theory such as the functional form of utility/profit functions, forms of strategic interaction, and any equilibrium or optimality concepts. It is important to note that G can be set-valued. This allows encoding the researcher's lack of understanding of some part of the structural model. Specifically, it does not require the knowledge of the *selection mechanism* according to which the observed outcome s is *selected* from $G(u|x;\theta)$. This formulation nests the classic setting in which the model is characterized by a *reduced* *form equation*:

$$Y = g(u|X;\theta), \tag{2.1}$$

for a function $g : U \times \mathcal{X} \times \Theta \to \mathcal{Y}$. This corresponds to the setting in which G is almost surely singleton-valued, i.e. $G(u|x;\theta) = \{g(u|x;\theta)\}, a.s.$ If this is the case, we say the model makes a *complete prediction*.

Throughout, we assume that u's law belongs to a parametric family $F = \{F_\theta, \theta \in \Theta\}$, where, for each θ, F_θ is a probability distribution on U. To keep notation concise, we use the same θ for parameters that show up in G and that index F_θ. Also, we focus on settings in which u is independent of X. However, the framework can be easily extended to settings where u is correlated with X, and the researcher specifies its conditional distribution $F_\theta(u|x)$. Furthermore, we note that our framework does accommodate settings in which some of the observable covariates are endogenous but one can construct a set-valued control function (See Example 3 below).

2.1 Motivating examples

Below, we illustrate the objects introduced above with examples studied in the literature. Our first two examples are discrete games of complete information (Bresnahan and Reiss, 1991a; Ciliberto and Tamer, 2009).

Example 1 (Discrete Games of Strategic Substitution)**.** There are two players (e.g. firms). Each player may either choose $y^{(j)} = 0$ or $y^{(j)} = 1$. The payoff of player j is

$$\pi^{(j)} = y^{(j)}\big(x^{(j)\prime}\delta^{(j)} + \beta^{(j)}y^{(-j)} + u^{(j)}\big), \tag{2.2}$$

where $y^{(-j)} \in \{0,1\}$ denotes the other player's action, $x^{(j)}$ is player j's observable characteristics of player j, and $u^{(j)}$ is an unobservable payoff shifter. The payoff is summarized below and is assumed to belong to the players' common knowledge.

		Player 2	
		$y^{(2)} = 0$	$y^{(2)} = 1$
Player 1	$y^{(1)} = 0$	$0, 0$	$0, x^{(2)\prime}\delta^{(2)} + u^{(2)}$
	$y^{(1)} = 1$	$x^{(1)\prime}\delta^{(1)} + u^{(1)}, 0$	$x^{(1)\prime}\delta^{(1)} + \beta^{(1)} + u^{(1)}$, $x^{(2)\prime}\delta^{(2)} + \beta^{(2)} + u^{(2)}$

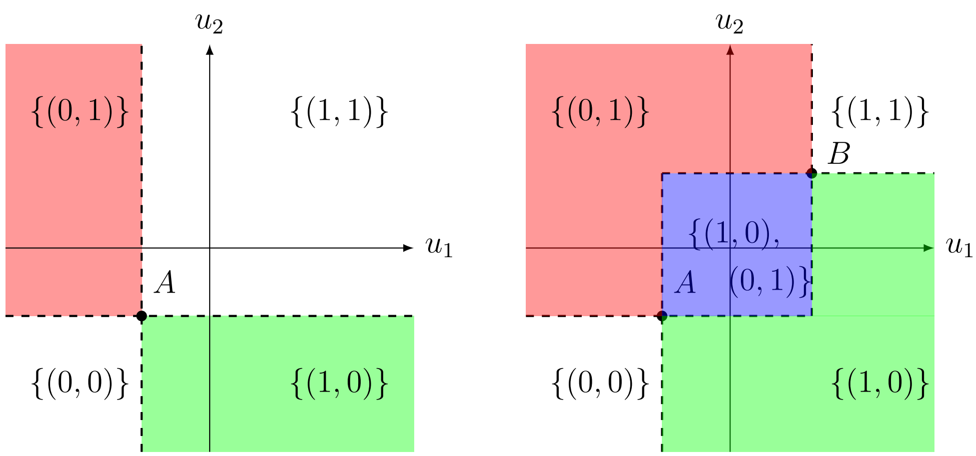
The key parameter is the *strategic interaction effect* $\beta^{(j)}$ which captures the impact of the opponent's taking $y^{(-j)} = 1$ on player j's payoff. Suppose that $\beta^{(j)} \leq 0$ for both players. For example, if the outcome represents each firm's market entry, $\beta^{(j)}$ measures the effect of the other firm's entry on firm j's profit. Let $\theta = (\beta, \delta)$. Suppose that the players play a pure strategy Nash equilibrium (PSNE). Then, the set of PSNEs predicted by this model is summarized by the following map:

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}, \\ \{(0,1)\} & u \in U_2, \\ \{(1,0)\} & u \in U_1, \\ \{(1,1)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}, \\ \{(1,0),(0,1)\} & -x^{(j)\prime}\delta^{(j)} < u^{(j)} < -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, \quad j = 1,2. \end{cases} \tag{2.3}$$

where $U_1 = \{u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$ and $U_2 = \{u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} > x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}$.

Figure 1 shows the level sets of $u \mapsto G(u|x,\theta)$ for a given (x,θ). When $\beta^{(j)} < 0$ for both players, the model admits multiple equilibria $\{(0,1),(1,0)\}$ when each $u^{(j)}$ is between the two thresholds $x^{(j)\prime}\delta^{(j)}$ and $x^{(j)\prime}\delta^{(j)} - \beta^{(j)}$ (the blue region in Figure 1). When $\beta^{(j)} = 0$ for either of the players, the model predicts a unique equilibrium for any value of $u = (u^{(1)}, u^{(2)})'$ (see left panel of Figure 1, in which $\beta^{(j)} = 0, j = 1,2$).

Figure 1: Level sets of $u \mapsto G(u|x;\theta)$



Note: $A = (-x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)})$; $B = (-x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)})$.
Left Panel: $\beta^{(1)} = \beta^{(2)} = 0$ and the model is complete. Right panel: $\beta^{(1)} < 0$ and $\beta^{(2)} < 0$ and the model is incomplete. U_1 in (2.3) corresponds to the region in green, and similarly U_2 is the region in red. Multiple equiribria $\{(1,0),(0,1)\}$ are predicted in the blue region.

Example 2 (Discrete Games of Strategic Complementarity)**.** Consider the payoff functions in (2.2) again but assume that $\beta^{(j)} \geq 0$. This setting can be used to analyze households' labor supply or retirement decisions, in which household members' labor force participation can be strategically complementary (Bresnahan and Reiss, 1991a).

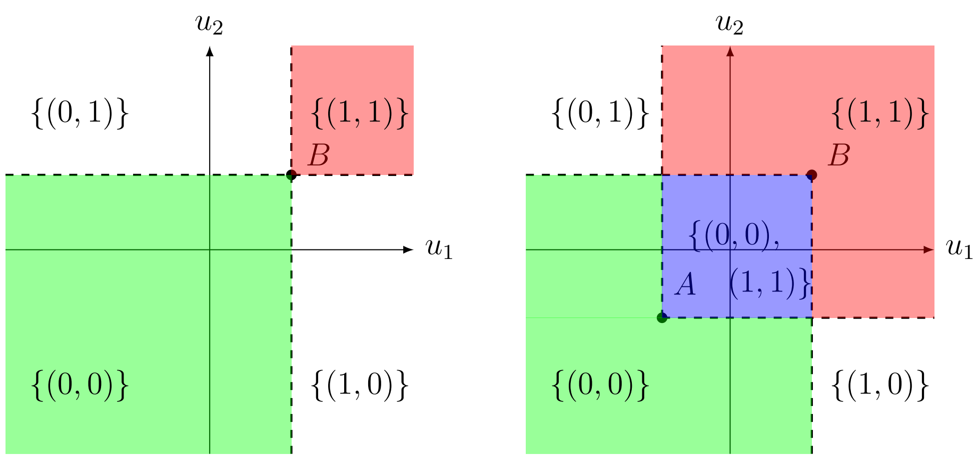
For each (x, u), the predicted set of PSNE's is given by

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in U_1, \\ \{(0,1)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \geq -x^{(2)\prime}\delta^{(2)} \\ \{(1,0)\} & u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \\ \{(1,1)\} & u \in U_2, \\ \{(0,0),(1,1)\} & -x^{(j)\prime}\delta^{(j)} - \beta^{(j)} \leq u^{(j)} < -x^{(j)\prime}\delta^{(j)},\ j = 1,2, \end{cases} \tag{2.4}$$

where $U_1 = \{u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} - \beta^{(1)} \leq u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$, and $U_2 = \{u^{(1)} \geq -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \geq -x^{(2)\prime}\delta^{(2)}\} \cup \{u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \leq u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$.

When $\beta^{(j)} = 0$ for one of the players, the model makes a complete prediction for almost all u. In contrast, if $\beta^{(j)} > 0$ for both members, both $(0,0)$ and $(1,1)$ can arise as equilibrium outcomes for some value of u.

Figure 2: Level sets of $u \mapsto G(u|x;\theta)$



Note: $A = (-x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)})$; $B = (-x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)})$.
Left Panel: $\beta^{(1)} = \beta^{(2)} = 0$ and the model is complete. Right panel: $\beta^{(1)} > 0$ and $\beta^{(2)} > 0$ and the model is incomplete. U_1 in (2.4) corresponds to the region in green, and similarly U_2 is the region in red. Multiple equiribria $\{(0,0),(1,1)\}$ are predicted in the blue region.

The next example is a parametric version of the triangular system of nonseparable equations (Chesher, 2003; Shaikh and Vytlacil, 2011). We consider a control function approach to this model.

Example 3 (Triangular model with an incomplete control function)**.** Consider a triangular model, in which a binary outcome y_i is determined by a binary treatment d_i, a vector w of exogenous covariates, and an unobserved variable ϵ_i; the treatment indicator d_i is determined by a vector of instrumental variables z_i and an unobserved variable v_i:

$$y_i = 1\{\alpha d_i + w_i'\eta + \epsilon_i \geq 0\}, \tag{2.5}$$
$$d_i = 1\{z_i'\gamma + v_i \geq 0\}. \tag{2.6}$$

Suppose that (w_i, z_i) is independent of (ϵ_i, v_i). The unobserved characteristics ϵ_i and v_i may be dependent, making d_i potentially endogenous.

If one could recover the unobservable characteristic v_i from the observables (which would be possible with a continuous d_i), conditioning on v_i would make ϵ_i independent of d_i. This *control function approach* would allow us to recover key model parameters (Imbens and Newey, 2009; Wooldridge, 2015). In the current setting, we may not uniquely recover v_i due to the discreteness of d_i, which makes it difficult to apply the control function approach without further assumptions.[1] However, the model restricts v_i to the following set:

$$\begin{aligned} H(d_i, z_i; \gamma) &\equiv \big\{v \in \mathbb{R} : d_i = 1\{z_i'\gamma + v \geq 0\}\big\} \\ &= \begin{cases} [-z_i'\gamma, \infty) & \text{if } d_i = 1 \\ (-\infty, -z_i'\gamma) & \text{if } d_i = 0. \end{cases} \end{aligned} \tag{2.7}$$

Suppose that ϵ_i's conditional distribution given v_i belongs to a location family and the location parameter is βv_i. Then, one may write $\epsilon_i = \beta v_i + u_i$ for some u_i independent of d_i. Substituting this expression into (2.5) and noting that $v_i \in H(d_i, z_i; \gamma)$, the set of outcome values compatible with the model is

$$G(u_i|x_i;\theta) = \Big\{y_i \in \{0,1\} : y_i = 1\{\alpha d_i + w_i'\eta + \beta v_i + u_i \geq 0\},\ \text{for some } v_i \in H(d_i, z_i; \gamma)\Big\}, \tag{2.8}$$

where $x_i = (d_i, w_i', z_i')'$ and $\theta = (\beta, \delta')'$ with $\delta = (\alpha, \eta', \gamma')'$. One of the benefits of the control function approach that one can test the endogeneity of d_i (Wooldridge, 2015). As we show below, this is also the case even if the control function cannot be uniquely recovered.[2]

The next example is a panel dynamic discrete choice model (Heckman, 1978; Hyslop, 1999).

[1] Wooldridge (2014) uses the generalized residual $r_i = d_i\lambda(z_i'\gamma) - (1-d_i)\lambda(-z_i'\gamma)$ from the first stage MLE, where λ is the inverse Mills ratio. He makes additional high-level assumptions so that r_i serves as a sufficient statistic for capturing the endogeneity of d_i and proposes an estimator of the average structural function. Instead of taking this approach, we explore what can be learned from the set-valued control function.

[2] We take a control function approach that conditions on v_i, which only requires specification of the conditional distribution of ϵ_i given v_i. Alternatively, one could specify the joint distribution of (ϵ_i, v_i). This alternative but stronger assumption would imply a complete model; there is a unique value of (y_i, d_i) for a given (ϵ_i, v_i) and exogenous covariates due to the triangular structure (Lewbel, 2007).

Example 4 (Panel Dynamic Discrete Choice Models)**.** An individual makes binary decisions across multiple periods according to

$$y_{it} = 1\{x_{it}'\lambda + y_{it-1}\beta + \alpha_i + \epsilon_{it} \geq 0\},\ i = 1, \dots, n,\ t = 1, \dots, T, \tag{2.9}$$

where y_{it} is a binary outcome for individual i in period t, x_{it} is a vector of observable covariates, α_i is an unobservable individual specific effect, and ϵ_{it} is an unobserved idiosyncratic error. If β is nonzero, the individual's choice in period t depends on her past choice, rendering the decision *state dependent*.

Suppose the researcher observes (y_{it}, x_{it}) for $i = 1, \dots, n$ and $t = 1, \dots, T$. Since y_{i0} is not observable, this leaves the value of $y_{i1}, \dots, y_{iT}$ not fully determined and makes the model incomplete (Heckman, 1978, 1987; Honoré and Tamer, 2006).[3] For example, consider $T = 2$. Suppose for the moment $y_{i0} = 0$. For a given $(x_i, \alpha_i, \epsilon_{i1}, \epsilon_{i2})$, the observed outcome $y_i = (y_{i1}, y_{i2})$ must satisfy

$$y_{i1} = 1\{x_{i1}'\lambda + \alpha_i + \epsilon_{i1} \geq 0\} \tag{2.10}$$

$$y_{i2} = 1\{x_{i2}'\lambda + y_{i1}\beta + \alpha_i + \epsilon_{i2} \geq 0\}. \tag{2.11}$$

Similarly, if $y_{i0} = 1$, the outcome must satisfy

$$y_{i1} = 1\{x_{i1}'\lambda + \beta + \alpha_i + \epsilon_{i1} \geq 0\} \tag{2.12}$$

$$y_{i2} = 1\{x_{i2}'\lambda + y_{i1}\beta + \alpha_i + \epsilon_{i2} \geq 0\}. \tag{2.13}$$

Without further assumptions, the model permits both possibilities. Letting $u_i = (u_{i1}, u_{i2})'$ with $u_{it} = \alpha_i + \epsilon_{it}$, the model prediction can therefore be summarized by the following correspondence

$$G(u_i|x_i;\theta) = \Big\{y_i = (y_{i1}, y_{i2}) \in \{0,1\}^2 : y_i \text{ satisfies either (2.10)-(2.11) or (2.12)-(2.13)}\Big\}. \tag{2.14}$$

If $\beta \geq 0$, this map can be expressed as follows:[4]

$$G(u_i|x_i;\theta) = \begin{cases} \{(0,0)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda, \\ \{(0,1)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda, \\ \{(1,0)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda - \beta, \\ \{(1,1)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,0)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \leq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ -x_{i2}'\lambda - \beta \leq u_{i2} < -x_{i2}'\lambda, \\ \{(0,1),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda. \end{cases} \tag{2.15}$$

[3] As an alternative, one could work with the likelihood function conditional on the initial observation. However, this approach can be problematic if one wants to be internally consistent across a different number of periods (Honoré and Tamer, 2006; Wooldridge, 2005).

[4] Appendix B.3 provides details and a graphical illustration of G.

Similar to the previous examples, the model makes a complete prediction when $\beta = 0$ (see Figure 6 in the Appendix).

3 Testing Hypotheses

Let $\beta \in \Theta_\beta \subset \mathbb{R}^{d_\beta}$ denote the subvector of θ whose value determines whether the model is complete or not. Let $\delta \in \Theta_\delta \subset \mathbb{R}^{d_\delta}$ collect the remaining components of θ. Given a sample of data $(Y_i, X_i), i = 1, \ldots, n$, consider testing a hypothesis on β. Let the null and alternative hypotheses be

$$H_0 : \beta = \beta_0, \quad \text{v.s.} \quad H_1 : \beta \in B_1, \tag{3.1}$$

where $B_1 \subset \Theta_\beta$ is a set not containing β_0. For instance, in entry games (i.e. Example 1), the presence of strategic substitution effects can be tested by letting $\beta_0 = 0$ and $B_1 = \{\beta : \beta^{(j)} < 0, j = 1, 2\}$. Similarly, we may test the potential endogeneity of treatment assignments (Example 3) and the presence of state dependence (Example 4) by setting $\beta_0 = 0$ and choosing suitable alternative hypotheses. In what follows, we let $\Theta_0 = \{\beta_0\} \times \Theta_\delta$ and $\Theta_1 = B_1 \times \Theta_\delta$ denote the sets of null and alternative parameter values respectively.

Let $\Delta_{Y|X}$ denote the set of conditional distributions (or probability kernels) of Y given $X = x$. For each $\theta = (\beta', \delta')'$, an incomplete model admits the following set of conditional distributions:

$$\mathcal{Q}_\theta = \Big\{ Q \in \Delta_{Y|X} : Q(A|x) = \int_U p(A|x,u) dF_\theta(u), \ \forall A \subseteq \mathcal{Y}, \\ \text{for some } p \in \Delta_{Y|X,u} \text{ such that } p(G(u|x;\theta)|x,u) = 1, \ a.s. \Big\}. \tag{3.2}$$

Here, the conditional distribution $p(\cdot|x, u)$ represents the unknown *selection mechanism* according to which an outcome gets selected from the set of predicted outcome values. Since the model is silent about its specification, we allow any law supported on $G(u|x;\theta)$. This means that the model can admit (infinitely) many likelihood functions for a given θ. Let μ be the counting measure on $\mathcal{Y}$. For each θ, define

$$\mathfrak{q}_\theta = \{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu, Q \in \mathcal{Q}_\theta\}. \tag{3.3}$$

This set collects all (conditional) densities that are compatible with a given θ. In the case of discrete games (Examples 1 and 2), this set collects all densities of equilibrium outcomes that are compatible with the description of the game. Similarly, in the context of panel discrete choice (Example 4), this set collects all densities of individual choices compatible with arbitrary specifications of the initial condition. The multiplicity of the densities is due to the model incompleteness that admits any selection mechanism $p(\cdot|u, x)$. In this sense, we may think of elements in $\mathfrak{q}_\theta$ being implicitly indexed by the unknown selection mechanisms. Observe that

$\mathfrak{q}_\theta$ reduces to a singleton set $\{q_\theta\}$ if the model is complete, i.e. $G(u|x;\theta) = \{g(u|x;\theta)\}$ for some function g, in which case $q_\theta = dQ_\theta/d\mu$ with $Q_\theta(A|x) = \int 1\{g(u|x;\theta) \in A\}dF_\theta(u)$.

While the multiplicity of likelihood functions may appear challenging, $\mathfrak{q}_\theta$ can be simplified, and this property enables us to conduct robust tests in a tractable manner. By Artstein's inequality (see e.g. Galichon and Henry, 2011; Molinari, 2020), $\mathfrak{q}_\theta$ can be written as the following set of densities satisfying a finite number of linear inequalities :

$$\mathfrak{q}_\theta = \Big\{q_{y|x} : \sum_{y\in A} q_{y|x}(y|x) \geq \nu_\theta(A|x),\ A \subseteq \mathcal{Y}\Big\}, \tag{3.4}$$

where

$$\nu_\theta(\cdot|x) = F_\theta(G(u|x;\theta) \subseteq \cdot|x) \tag{3.5}$$

is the conditional *containment functional* (or *belief function*) associated with the random set $G(u|x;\theta)$. This functional gives the sharp lower bound for the conditional probability $Q(A|x)$ across all Q's belonging to $\mathcal{Q}_\theta$.[5] Theoretical properties of $\nu_\theta(\cdot|x)$ and numerical methods for computing it are well studied in the literature (Ciliberto and Tamer, 2009; Galichon and Henry, 2011).[6] For us, the fact that $\mathfrak{q}_\theta$ is characterized by a system of linear inequalities is important. Together with an extended Neyman-Pearson lemma reviewed below, this allows us to construct a computationally tractable score-based test. In the next subsection, we briefly review the existing results we will rely on.

3.1 Preliminaries

Let $p_0(y|x)$ denote the true conditional distribution of the outcome given the covariates. Let us start with a problem of distinguishing $\theta = \theta_0$ from another value $\theta = \theta_1$ with $\theta_0 \neq \theta_1$. In a parametrically specified complete model $\{p_\theta, \theta \in \Theta\}$, this amounts to testing $p_0 = p_{\theta_0}$ against $p_0 = p_{\theta_1}$. It is well known that the most powerful test for such a problem is a likelihood-ratio test, which is the consequence of the Neyman-Pearson lemma. In incomplete models, corresponding null and alternative hypotheses would be $p_0 \in \mathfrak{q}_{\theta_0}$ and $p_0 \in \mathfrak{q}_{\theta_1}$ rendering both hypotheses composite. KZ19 show that it is possible to extend the Neyman-Pearson lemma to such settings, building on a general result established by Huber and Strassen (1973). Their key observation is that there is a *least favorable pair (LFP)* $(q_{\theta_0}, q_{\theta_1}) \in \mathfrak{q}_{\theta_0} \times \mathfrak{q}_{\theta_1}$ of densities. This pair is such that q_{θ_0} is consistent with θ_0 and is least favorable for controlling the size of a test among all densities belonging to $\mathfrak{q}_{\theta_0}$, whereas q_{θ_1} is consistent with θ_1 and is least favorable for maximizing a measure of power among the densities belonging to $\mathfrak{q}_{\theta_1}$.[7]

[5] The upper bound for $Q(A|x)$ is given by the capacity functional $\nu^*(A|X) = F_\theta(G(u|x;\theta) \cap A \neq \varnothing|x)$ (Molinari, 2020). It is sufficient to use either of the lower or upper bounds in (3.4) because the bounds are related to each other through the conjugate relationship $\nu(A|x) = 1 - \nu^*(A^c|x)$.

[6] We provide a brief review on these in Appendix A.

[7] They consider the lower envelope of power over $\mathcal{Q}_\theta$.

Furthermore, they show that an LR-test based on this pair constitutes a minmax test, which maximizes a robust measure of power among a class of level-α tests (see KZ19 Section 3).

There is a simple way to compute the LFP through a convex program. For each $x \in \mathcal{X}$, the LFP $(q_{\theta_0}, q_{\theta_1})$ is characterized as

$$(q_{\theta_0}, q_{\theta_1}) = \arg\min_{(q_0,q_1)} \sum_{y\in\mathcal{Y}} \ln\Big(\frac{q_0(y|x)+q_1(y|x)}{q_0(y|x)}\Big)(q_0(y|x)+q_1(y|x)) \tag{3.6}$$

$$s.t. \sum_{y\in A} q_0(y|x) \geq \nu_{\theta_0}(A|x),\ \ A \subseteq \mathcal{Y} \tag{3.7}$$

$$\sum_{y\in A} q_1(y|x) \geq \nu_{\theta_1}(A|x),\ \ A \subseteq \mathcal{Y}. \tag{3.8}$$

The constraints in (3.7) and (3.8) are the sharp identifying restrictions.[8] In view of (3.4), they are equivalent to saying that q_0 belongs to $\mathfrak{q}_{\theta_0}$ and q_1 belongs to $\mathfrak{q}_{\theta_1}$ respectively. For us, these restrictions are useful for computing the LFP because they are linear in (q_0, q_1). The convex problem can be solved numerically in general. In some of the leading examples, it is also possible to compute it analytically.

To illustrate, let us consider Example 1. Suppose that the latent payoff shifters $(u^{(1)}, u^{(2)})$ follow a bivariate standard normal distribution. We may then compute $\nu_\theta(A|x)$ for each event. Let us take $A = \{(1,0)\}$ as an example. Using (2.3) and (3.5), we obtain

$$\begin{aligned}\nu_\theta(\{(1,0)\}|x) &= F_\theta(G(u|x;\theta) \subseteq \{(1,0)\}|x)\\ &= F_\theta(u \in U_1) = (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})].\end{aligned} \tag{3.9}$$

This corresponds to the probability assigned to the green region in Figure 1 (right panel) and is the sharp lower bound for the probability of $A = \{(1,0)\}$.

Now consider two parameter values $\theta_0 = (0_2', \delta')'$ and $\theta_1 = (\beta', \delta')'$, where $\beta = (\beta^{(1)}, \beta^{(2)})'$ with $\beta^{(j)} < 0$ for $j = 1,2$. As we discuss in more detail below, the model is complete when $\beta = 0$. One can show that the restrictions in (3.7) reduce to the following equality restrictions:

$$q_0((0,0)|x) = (1-\Phi(x^{(1)\prime}\delta^{(1)}))(1-\Phi(x^{(2)\prime}\delta^{(2)})) \tag{3.10}$$

$$q_0((0,1)|x) = (1-\Phi(x^{(1)\prime}\delta^{(1)}))\Phi(x^{(2)\prime}\delta^{(2)}) \tag{3.11}$$

$$q_0((1,0)|x) = \Phi(x^{(1)\prime}\delta^{(1)})(1-\Phi(x^{(2)\prime}\delta^{(2)})) \tag{3.12}$$

$$q_0((1,1)|x) = \Phi(x^{(1)\prime}\delta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}). \tag{3.13}$$

[8] A common way to use them for identification analysis is to define the *sharp identified set* as $\Theta_I = \{\theta : P(A|x) \geq \nu_\theta(A|x), a.s.\}$. That is, given the conditional probability $P(\cdot|x)$ identified from data, one collects all values of θ satisfying the sharp identifying restrictions. For hypothesis testing, we instead fix θ and ask what would be a distribution among all distributions satisfying the sharp identifying restrictions that is least favorable for controlling the size or maximizing the power.

These restrictions uniquely determine the least-favorable null density q_{θ_0}. Hence, we may write

$$q_{\theta_0}(y|x) = [(1-\Phi_1)(1-\Phi_2)]^{1\{y=(0,0)\}}[(1-\Phi_1)\Phi_2]^{1\{y=(0,1)\}} \\ \times [\Phi_1(1-\Phi_2)]^{1\{y=(1,0)\}}[\Phi_1\Phi_2]^{1\{y=(1,1)\}}, \quad (3.14)$$

where, to ease notation, we use Φ_1 and Φ_2 to denote $\Phi(x^{(1)\prime}\delta^{(1)})$ and $\Phi(x^{(2)\prime}\delta^{(2)})$.

When $\beta^{(j)} < 0, j = 1, 2$, there are multiple densities satisfying (3.8). The least favorable alternative density q_{θ_1} can be found by minimizing (3.6) with respect to q_1 subject to (3.8). The solution can be expressed analytically. For example, when player 1's strategic interaction effect on player 2 is relatively high, it is given by the following form:[9]

$$q_{\theta_1}(y|x) = [(1-\Phi_1)(1-\Phi_2)]^{1\{y=(0,0)\}}[(1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)}))\Phi_2]^{1\{y=(0,1)\}} \\ \times [(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})(\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))]^{1\{y=(1,0)\}} \\ \times [\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]^{1\{y=(1,1)\}}. \quad (3.15)$$

Comparing (3.14) and (3.15), one can see that q_{θ_1} tends to q_{θ_0} as β approaches its null value (i.e. 0). Hence, one may view $\theta \mapsto q_\theta$ a "parametric" model, and we will indeed take this view below. Here is a way to interpret this parametric model. For each θ_1, the density q_{θ_1} corresponds to the data generating process that is least favorable in terms of detecting β's deviation from its null value among all densities compatible with θ_1. Behind q_{θ_1}, there is a selection mechanism that induced the least favorable DGP. For our purposes, however, we do not need to know the precise form of this selection mechanism. When we solve the convex program, we are "profiling out" the selection mechanism and directly obtaining the induced density q_{θ_1}. This is why q_{θ_1} no longer involves any selection mechanism.

By varying θ_1, we may trace out a family of such densities and form a parametric model. We therefore call the map $\theta \to q_\theta$ the *least favorable (LF) parametric model.* Equation (3.15) suggests that we may pretend as if data were generated by a parametric discrete choice model with the given density. This is indeed the case if one is interested in maximizing a measure of discrimination between θ_0 and θ_1 based on q_θ. Thanks to this property, most of our analysis below will resemble that of standard discrete choice models, which helps us keep our framework tractable.

3.2 Model completeness under the null

The following assumption imposes a key structure on the model.

Assumption 1. *(i) For any null parameter value $\theta_0 = (\beta_0', \delta')'$ with $\delta \in \Theta_\delta$, the set of conditional densities of outcome is a singleton $\mathfrak{q}_{\theta_0} = \{q_{\theta_0}\}$; (ii) For any pair of parameters $\theta_0 = (\beta_0', \delta')'$ and $\theta_1 = (\beta', \delta')'$ with $\beta \in B_1$ and $\delta \in \Theta_\delta$, we have $\mathfrak{q}_{\theta_0} \cap \mathfrak{q}_{\theta_1} = \varnothing$.*

[9] See Appendix B.1 for details.

By Assumption 1 (i), we require the model is complete under the null hypothesis in the sense that $\mathfrak{q}_{\theta_0}$ contains a unique density when $\beta = \beta_0$. As discussed earlier, this holds whenever the model makes a complete prediction under the null hypothesis and is satisfied in the examples discussed in Section 2.1.

The model can be complete or incomplete under the alternative hypothesis. Assumption 1 (ii) requires that the sets $\mathfrak{q}_{\theta_0}$ and $\mathfrak{q}_{\theta_1}$ are disjoint. If this is the case, it is possible to detect θ_1's local deviation from θ_0 regardless of the unknown selection mechanism. In KZ19, such an alternative hypothesis is called *robustly testable*, and we focus on settings in which this assumption is satisfied.[10]

Let us now revisit the examples.

Example 1 (Binary Response Game of Complete Information)**.** Consider testing the presence of strategic substitution effects by testing $H_0 : \beta^{(1)} = \beta^{(2)} = 0$ against $H_0 : \beta^{(1)} < 0, \beta^{(2)} < 0$. Under the null hypothesis, there is no strategic interaction between the players, which leads to the following complete prediction:

$$G(u|x;\theta_0) = \begin{cases} \{(0,0)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}, \\ \{(1,1)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}, \\ \{(1,0)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)}, u^{(2)} \leq -x^{(2)\prime}\delta^{(2)}, \\ \{(0,1)\} & u^{(1)} \leq -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}. \end{cases} \tag{3.16}$$

Hence, for any value of the observed and unobserved variables, $G(u|x;\theta_0)$ contains a unique equilibrium outcome. This corresponds to the left panel of Figure 1. A similar analysis applies to Example 2.

Example 3 (Triangular Model with an Incomplete Control Function)**.** Consider testing the endogeneity of the treatment by testing the hypothesis that the coefficient β on the control function v is 0. When the null hypothesis is true, the model's prediction reduces to

$$y_i = 1\{\alpha d_i + w_i'\eta + u_i \geq 0\}. \tag{3.17}$$

Hence, for a given (x_i, u_i), the value of y_i is uniquely determined regardless of the value of the control function. Indeed, there is no need to control for v_i because u_i is independent of (d_i, w_i). Hence, this is a model of binary choice with exogenous covariates whose analysis is standard.

Example 4 (Panel Dynamic Discrete Choice Models)**.** Consider testing the presence of state dependence. This can be done by testing whether the coefficient β on the lagged dependent

[10] KZ19 analyze a general case in which this assumption may fail to hold by extending the notion of local alternatives. We conjecture that we may extend our framework similarly. Since our leading examples satisfy Assumption 1 (ii) (see Appendix B), we leave this extension elsewhere.

variable y_{it-1} is 0 or not. When $\beta = 0$ in (2.9), the model reduces to the static panel binary choice model

$$y_{it} = 1\{x_{it}'\lambda + \alpha_i + \epsilon_{it} \geq 0\},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{3.18}$$

which makes (2.10)-(2.11) and (2.12)-(2.13) equivalent. Hence, under the null hypothesis, $G(u_i|x_i;\theta_0)$ contains the unique outcome satisfying (3.18).

3.3 Score-based Tests

Score based tests such as Rao's score (or Lagrange multiplier) test and Neyman's $C(\alpha)$ test are widely used. These tests require estimation of the restricted model only, which is attractive in our setting. The restricted model is complete and hence typically admits point estimation of nuisance parameters under fairly weak conditions. We take advantage of this property to carry out a score-based test. Below, we briefly review key ideas behind the classic score tests and discuss extensions to handle potential model incompleteness under the alternative. For expositional purposes, we assume that q_θ is differentiable with respect to θ for now and will weaken this assumption later.

Consider testing the null parameter value $\theta_0 = (\beta_0', \delta')'$ against a local alternative hypothesis $\theta_h = (\beta_0' + h', \delta')'$, where $h \in \mathbb{R}^{d_\beta}$. The most powerful test for this problem is the likelihood-ratio test, which compares q_{θ_h} to q_{θ_0} and rejects H_0 when the ratio of the two is high (KZ19). The test is also robust in the sense that, under Assumption 1 (ii), the log-likelihood ratio can detect any deviation from the null hypothesis with non-trivial power no matter what the selection mechanism is. The log-likelihood ratio can be locally approximated by $\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta)$, where $s_\beta(y|x;\beta,\delta) = \frac{\partial}{\partial\beta}\ln q_\theta(y|x)|_{\theta=(\beta,\delta)}$ is the score function. Let $\Sigma_{\beta_0} = \mathrm{Var}(\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta))$. For a fixed h, the normalized quantity

$$\frac{(\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta))^2}{h'\Sigma_\beta h}, \tag{3.19}$$

serves as a measure of discrimination between β_0 and $\beta_0 + h$. This is a robust measure of local discrimination between β_0 and $\beta_0 + h$ because it approximates how the log-likelihood $\ln q_{\theta_h}$ changes from $\ln q_{\theta_0}$, which is shown to be the robust and optimal way to detect the local deviation from the null hypothesis (KZ19).

For i.i.d. data, $\Sigma_{\beta_0} = nI_{\beta_0}$ where $I_{\beta_0} = E[s_\beta(Y_i|X_i;\beta_0,\delta)s_\beta(Y_i|X_i;\beta_0,\delta)']$. If one seeks for a direction h that maximizes (3.19), it is given by $h^* = I_{\beta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta)$, which motivates Rao's score statistic:[11]

$$T_n = \sup_{h\in\mathbb{R}^{d_\beta}} \frac{\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta))^2}{nh'I_{\beta_0}h} = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta)' I_{\beta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta). \tag{3.20}$$

[11]See Bera and Bilias (2001) for a more detailed argument for complete models. The same argument can be applied to incomplete models by replacing the standard likelihood function with the LF density q_θ.

This statistic depends on the unknown nuisance parameter δ. Suppose that the nuisance parameter δ can be estimated by a point estimator $\hat{\delta}_n$. Evaluating the sample mean of the score at $\delta = \hat{\delta}_n$ and imposing the null hypothesis yields

$$g_n(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n). \tag{3.21}$$

A feasible version of (3.20) is

$$\hat{T}_n = g_n(\beta_0)'\hat{V}_n^{-1} g_n(\beta_0), \tag{3.22}$$

where $\hat{V}_n = n^{-1}\sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n)s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n)'$ is a consistent estimator of the asymptotic variance $V_0 \equiv I_{\beta_0}$. The sampling distribution of the score generally depends on $\hat{\delta}_n$. However, if one uses the restricted MLE (discussed in the next section) to estimate δ, which we recommend, $\hat{T}_n$ converges in distribution to a χ^2-distribution with d_β degrees of freedom under the null hypothesis. It is also possible to use point estimators other than the restricted MLE, in which case we recommend using an orthogonalized version of the score in the spirit of Neyman's $C(\alpha)$ test (see Remark 1 below).

The analysis so far presumed that q_θ was differentiable and $h \in \mathbb{R}^{d_\beta}$ was unrestricted. These assumptions may be too restrictive in some settings. For example, in discrete games of complete information, the least favorable parametric model $h \mapsto q_{\theta_h}$ and its score take different functional forms depending on whether the alternative hypothesis admits strategic substitution (i.e. $h < 0$ as in Example 1) or strategic complementarity (i.e. $h > 0$ as in Example 2). It is then natural to analyze these two cases separately. Below, we, therefore, weaken differentiability requirements to accommodate these features and also allow restrictions on the alternative hypothesis.

Let $\mathbb{C}(0,\epsilon)$ denote an open cube centered at the origin with edges of length 2ϵ. A set $\Gamma \subseteq \mathbb{R}^d$ is said to be locally equal to set $\Upsilon \subseteq \mathbb{R}^d$ if $\Gamma \cap \mathbb{C}(0,\epsilon) = \Upsilon \cap \mathbb{C}(0,\epsilon)$ for some $\epsilon > 0$ (Andrews, 1999).

Assumption 2 (L^2-directional differentiability)**.** *(i) $B_1 - \beta_0$ is locally equal to a convex cone $\mathcal{V}_1$; (ii) For any $\zeta \in \mathcal{V}_1 \times \mathbb{R}^{d_\delta}$, there exists a square integrable function $s_\theta = (s_\beta', s_\delta')' : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^d$ such that*

$$\left\| q_{\theta_0+\tau\zeta}^{1/2} - q_{\theta_0}^{1/2}\left(1 + \frac{1}{2}\tau\zeta' s_\theta(\cdot|\cdot;\beta_0,\delta)\right)\right\|_{L^2_\mu} = o(\tau), \tag{3.23}$$

as $\tau \downarrow 0$.

Assumption 2 (i) requires the set of deviations from β_0 can be locally approximated by a convex cone. In Example 1, consider testing $H_0 : \beta = (0,0)'$ against $H_1 : \beta^{(1)} < 0, \beta^{(2)} < 0$. Then, $B_1 - \beta_0$ is locally equal to

$$\mathcal{V}_1 = \{h = (h^{(1)}, h^{(2)}) : h^{(1)} < 0, h^{(2)} < 0\}. \tag{3.24}$$

Assumption 2 (ii) is a version of the differentiability in quadratic mean commonly used in the literature (see e.g van der Vaart, 2000). It only requires that a unique score, in the sense of the L^2-derivative of the square-root density, exists for the set $\mathcal{V}_1$ of local deviations from the null hypothesis, but it does not restrict the likelihood function otherwise. This weaker assumption is appropriate for incomplete models, and s_θ can be derived from the least favorable parametric model similar to the standard parametric models.[12]

Following Silvapulle and Silvapulle (1995), we now define a test statistic for $H_0 : \beta = \beta_0$ v.s. $H_1 : \beta \in B_1$ by

$$\hat{S}_n = g_n(\beta_0)'\hat{V}_n^{-1}g_n(\beta_0) - \inf_{h\in\mathcal{V}_1}(g_n(\beta_0) - h)'\hat{V}_n^{-1}(g_n(\beta_0) - h). \tag{3.25}$$

This test statistic is a slight modification of (3.22). It requires the same functions of data as T_n, but it is designed to direct power against the local alternatives in $\mathcal{V}_1$. While the asymptotic distribution of the statistic is no longer a χ^2 distribution, its critical value is easy to compute using simulations. Let

$$c_\alpha = \inf\{x \in \mathbb{R} : Pr(S \leq x) \geq 1 - \alpha\}, \tag{3.26}$$

where

$$S \equiv Z'V_0^{-1}Z - \inf_{h\in\mathcal{V}_1}(Z - h)'V_0^{-1}(Z - h),\ Z \sim N(0, V_0), \tag{3.27}$$

which can be simulated by drawing Z repeatedly from a zero mean multivariate normal distribution with estimated variance $\hat{V}_n$.

Remark 1. If one uses $\hat{\delta}_n$ other than the restricted MLE, $\hat{g}_n(\beta_0)$'s limiting distribution may depend on that of $\hat{\delta}_n$ in general. Neyman's $C(\alpha)$ statistic addresses this issue by making the statistic insensitive to the estimation error associated with $\hat{\delta}_n$. This is achieved by projecting s_β to s_δ and replacing g_n with the "orthogonalized" (or "residualized") score:

$$g_n(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n) - I_{\beta,\delta}I_\delta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s_\delta(Y_i|X_i;\beta_0,\hat{\delta}_n), \tag{3.28}$$

where $I_{\beta,\delta}$ and I_δ are submatrices of

$$I_\theta = \begin{bmatrix} I_\beta & I_{\beta,\delta} \\ I_{\delta,\beta} & I_\delta \end{bmatrix} = E_{q_\theta}\left[s_\theta(Y_i|X_i)s_\theta(Y_i|X_i)'\right], \tag{3.29}$$

which can be estimated by their sample analogs. The orthogonalized score $g_n(\beta_0)$ constructed this way is robust to the estimation error of δ, and its asymptotic distribution coincides with a version of the statistic which replaces $\hat{\delta}_n$ with the true value δ_0. The asymptotic variance of $g_n(\beta_0)$ is $V_0 = I_\beta - I_{\beta,\delta}I_\delta^{-1}I_{\delta,\beta}$. The test statistic in (3.25) can be constructed in the same way using g_n in (3.28) and a consistent estimator $\hat{V}_n$ of V_0. The way to calculate the critical value remains the same.[13]

[12] Appendix B derives s_θ for some of the examples.

[13] When $\hat{\delta}_n$ is the restricted maximum likelihood estimator, the second term on the right hand side of (3.28) becomes asymptotically negligible making $\hat{S}_n$ asymptotically equivalent to the version without the Neyman orthogonalization (Kocherlakota and Kocherlakota, 1991).

3.4 Estimation of nuisance parameters

Our tests require an estimator $\hat{\delta}_n$ of the nuisance parameter. A natural estimator of δ is the restricted maximum likelihood estimator (MLE) $\hat{\delta}_n$, which is a maximizer of

$$\mathbb{M}_n(\delta) \equiv \frac{1}{n}\sum_{i=1}^n \ln q_{\beta_0,\delta}(Y_i|X_i), \tag{3.30}$$

where $q_{\beta_0,\delta}$ is the conditional density of Y_i when $\beta = \beta_0$. Under our assumptions, this density also coincides with the least favorable parametric model evaluated at $\beta = \beta_0$. We therefore call the map $\delta \mapsto \ln q_{\beta_0,\delta}$ the *restricted log-likelihood function.* The complete model (under H_0) is often a standard discrete choice problem. As such, existing package software can be used to compute $\hat{\delta}_n$.

Example 1 (Binary Response Game (continued))**.** Under $H_0 : \beta^{(1)} = \beta^{(2)} = 0$, the model has a unique likelihood function as discussed in Section 3.1. The (restricted) maximum likelihood estimator $\hat{\delta}_n$ maximizes

$$\mathbb{M}_n(\delta) = \sum_{i=1}^n \Big(\ln 1\{y_i = (0,0)\}(1-\Phi_1)(1-\Phi_2) + \ln 1\{y_i = (0,1)\}(1-\Phi_1)\Phi_2$$
$$+ \ln 1\{y_i = (1,0)\}\Phi_1(1-\Phi_2) + \ln 1\{y_i = (1,1)\}\Phi_1\Phi_2 \Big),$$

where $\Phi_j = \Phi(x^{(j)\prime}\delta^{(j)}), j = 1, 2$. Alternatively, one can estimate δ by only using features of the model that are uniquely predicted. This strategy used earlier in the literature would maximize a likelihood function based on the empirical frequency of no entry ($y_i = (0,0)$), monopoly ($y_i = (0,1)$ or $(1,0)$), and duopoly ($y_i = (1,1)$) (Bresnahan and Reiss, 1991b; Berry, 1992).[14] If this estimator is used, we recommend using the orthogonalized score to construct the test statistic.

Example 3 (Triangular Model with an Incomplete Control Function)**.** Recall that, when $\beta = 0$, the model reduces to a binary choice model with exogenous covariates. If we assume $u_i \sim N(0,1)$, the conditional probability of $y_i = 1$ is

$$q_{\theta_0}(1|d_i, w_i, z_i) = \Phi(\alpha d_i + w_i'\eta), \tag{3.31}$$

which can be used to estimate $\delta = (\alpha, \eta)$ by the restricted MLE. This can be done by any software that may estimate probit models.

Example 4. A random effects probit model assumes α_i is independent of x_i and follows $N(0, \gamma^2)$, and $\epsilon_{i1}, \ldots, \epsilon_{iT}$ are independent standard normal random variables. This yields the following conditional density function for each i:

$$q_{\beta_0,\delta}(y_i|x_i) = \int \prod_{t=1}^T \Phi\big[(2y_{it} - 1)(x_{it}'\lambda + \gamma a)\big]\phi(a)da, \tag{3.32}$$

[14] If this alternative estimator is used, T_n is not asymptotically equivalent to Rao's score statistic in general.

which can be used to estimate $\delta = (\lambda, \gamma)$ using, for example, a simulated maximum likelihood estimator (Train, 2009).

3.5 Asymptotic Properties

We collect results on the asymptotic properties of our test. Throughout, we assume that $u^n = (u_1, \dots, u_n)$ is an independent and identically distributed (i.i.d.) sample drawn from F_θ, and $X^n = (X_1, \dots, X_n)$ is also an i.i.d. sample drawn from a distribution q_X^n. The joint distribution of the outcome sequence $Y^n = (Y_1, \dots, Y_n) \in \mathcal{Y}^n$ conditional on $x^n = (x_1, \dots, x_n)$ belongs to the following set:

$$\mathcal{Q}_\theta^n = \Big\{ Q : Q(A|x^n) = \int_{U^n} p(A|u^n, x^n) dF_\theta^n(u),\ \forall A \subseteq \mathcal{Y}^n, \\ \text{for some } p \in \Delta_{Y^n|X^n,u^n} \ \text{ such that } p(G^n(u^n|x^n;\theta)|u^n, x^n) = 1, a.s. \Big\}, \quad (3.33)$$

where F_θ^n denotes the joint law of u^n, and $G^n(u^n|x^n;\theta) = \prod_{i=1}^n G(u_i|x_i;\theta)$ is the Cartesian product of the set-valued predictions.[15] We then let $\mathcal{P}_\theta^n$ collect joint laws of (Y^n, X^n); each element P^n of $\mathcal{P}_\theta^n$ is such that the conditional law of Y^n given X^n belongs to $\mathcal{Q}_\theta^n$, and the law of X^n is q_X^n.

We start with conditions that ensure the $\sqrt{n}$-consistency of $\hat{\delta}_n$. They are mainly regularity conditions on the restricted log-likelihood function. Fixing β to its null value, one can view $q_{\beta_0,\delta}$ as the conditional density of y in a regular parametric model, in which δ is the only unknown parameter. As such, the conditions below parallel the ones in the literature.

Below, let $\delta_0 \in \Theta_\delta$ denote the true value of the nuisance component vector. For each $\delta \in \Theta_\delta$, let $\mathbb{M}(\delta) \equiv E[\ln q_{\beta_0,\delta}]$, where expectation is taken with respect to the conditional density q_{β_0,δ_0} and the distribution of X. Let $\mathbb{M}_n(\delta) \equiv n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i)$ be the sample counterpart of $\mathbb{M}$ and let $\mathbb{G}_n(\delta) \equiv \sqrt{n}(\mathbb{M}_n(\delta) - \mathbb{M}(\delta))$ be an empirical process indexed by δ.

Assumption 3. *(i-a) There is a continuous function $M : \Theta_\delta \to \mathbb{R}_+$ such that*

$$\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} |\ln q_{\beta_0,\delta}(y|x)| \le M(\delta);$$

(i-b) $\delta \mapsto \ln q_{\beta_0,\delta}(y|x)$ is Lipschitz continuous uniformly in (y, x). That is,

$$\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \big| \ln q_{\beta_0,\delta}(y|x) - \ln q_{\beta_0,\delta'}(y|x) \big| \lesssim \|\delta - \delta'\| \ \forall \delta, \delta' \in \Theta_\delta. \quad (3.34)$$

[15] Assuming u^n and X^n are i.i.d. does not imply Y^n is i.i.d. The set $\mathcal{Q}_\theta^n$ in general contains dependent and heterogeneous laws because the behavior of the selection mechanism across experiments is unrestricted (see Epstein et al., 2016). This does not create an issue for size properties of our test because $\mathcal{Q}_\theta^n$ reduces to a single i.i.d. law under the null hypothesis.

(i-c) $\delta \neq \delta_0 \Rightarrow q_{\beta_0,\delta}(y|x) \neq q_{\beta_0,\delta_0}(y|x)$ *with positive probability;*

(ii) Θ_δ *is a nonempty compact set; (iii)* $\hat{\delta}_n$ *is such that* $\mathbb{M}_n(\hat{\delta}_n) \geq \inf_{\delta \in \Theta_\delta} \mathbb{M}_n(\delta) + r_n$, *where, for any* h *and* $\epsilon > 0$, $\sup_{P^n \in \mathcal{P}^n_{\theta_0+h/\sqrt{n}}} P^n(|r_n| > \epsilon) \to 0$.

We also assume F_θ belongs a smooth parametric family in the following sense.

Assumption 4. *For each* $\theta \in \Theta$, F_θ *is absolutely continuous with respect to a* σ*-finite measure* ζ *on* U. *The Radon-Nikodym density* $f_\theta = dF_\theta/d\zeta$ *satisfies*

$$\|f_\theta - f_{\theta'}\|_{L^1_\zeta} \leq C\|\theta - \theta'\|, \ \forall \theta, \theta' \in \Theta, \tag{3.35}$$

for some $C > 0$.

Finally, the following condition requires that the population objective function is locally well behaved so that its value is informative about δ_0, and the supremum of an empirical log-likelihood process can be controlled over a neighborhood of δ_0 (see van der Vaart and Wellner, 1996, Sec. 3.2.2).

Assumption 5. *For every* δ *in a neighborhood of* δ_0,

$$\mathbb{M}(\delta) - \mathbb{M}(\delta_0) \lesssim -\|\delta - \delta_0\|^2. \tag{3.36}$$

Furthermore,

$$\sup_{P^n \in \mathcal{P}^n_{\theta_0+h/\sqrt{n}}} E^*_{P^n} \sup_{\delta \in B_\zeta(\delta_0)} |\mathbb{G}_n(\delta) - \mathbb{G}_n(\delta_0)| \lesssim \zeta, \tag{3.37}$$

where $B_\zeta(\delta_0) = \{\delta : \|\delta - \delta_0\| < \zeta\}$.

Under thse assumptions, the restricted MLE $\hat{\delta}_n$ is $\sqrt{n}$-consistent.

Proposition 3.1. *Suppose Assumptions 1-5 hold. Then,*

$$\sqrt{n}\|\hat{\delta}_n - \delta_0\| = O_{P^n}(1), \tag{3.38}$$

uniformly in $P^n \in \mathcal{P}^n_{\theta_0+h/\sqrt{n}}$.

Below, let $P^n_0 \in \mathcal{P}^n_{\theta_0}$ be the unique joint law of (Y^n, X^n) under the null hypothesis. Let $s_{\theta,j}$ be the j-th component of s_θ. Define

$$\Xi = \big\{\xi : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | \xi(y,x) = s_{\theta,j}(y|x;\beta_0,\delta)s_{\theta,k}(y|x;\beta_0,\delta), \ 1 \leq j,k \leq d, \delta \in \Theta_\delta\big\}. \tag{3.39}$$

We assume the elements of Ξ obey a uniform law of large numbers, i.e. Ξ is a Glivenko-Cantelli class.

Assumption 6.

$$\sup_{\xi\in\Xi}\Big|\frac{1}{n}\sum_{i=1}^{n}\xi(Y_i,X_i)-E_{P_0}[\xi(Y_i,X_i)]\Big|=o_{P_0^n}(1). \tag{3.40}$$

Suppose $\hat{S}_n$ as defined in (3.25) or it is constructed from the orthogonalized score using an $\sqrt{n}$-consistent estimator that is not necessarily the restricted MLE. The following theorem shows that the test controls its size.

Theorem 3.1. *Suppose Assumptions 1-6 hold. Let c_α be defined by (3.26). Then, for any $\alpha\in(0,1)$,*

$$\lim_{n\to\infty}P_0^n(\hat{S}_n>c_\alpha)=\alpha. \tag{3.41}$$

3.6 Inference on parameters

In some applications, the ultimate goal may be to make inference on the underlying parameter, for example, to construct confidence intervals for components of θ. While this is not our focus, we discuss a possible way to achieve this and leave its formal analysis to future work.

Consider constructing confidence intervals for a component or linear combination $p'\delta_0$ of δ_0.[16] According to Proposition 3.1, $\hat{\delta}_n$ is a $\sqrt{n}$-consistent estimator of δ_0 as long as the true value of β is in a neighborhood of β_0 whose radius is of order $n^{-1/2}$. It would be natural to use such an estimator to construct a confidence interval for δ_0 if the complete model is selected. A well-known challenge for such post-model selection inference is that a naive asymptotic approximation that disregards the model selection step may not be valid uniformly over a large class of data generating processes. Given this, we consider the following hybrid method.

Step 1: Compute S_n and $c_n=(\kappa_n\wedge 1)c_\alpha$, where κ_n is a sequence that tends to 0 slowly, e.g. $\kappa_n=(\ln n)^{-1/2}$;

Step 2:

- Reject $H_0:\beta=\beta_0$ if $S_n>c_n$. Construct a *robust confidence interval* by Kaido et al. (2019) for $p'\delta$;
- Do not reject $H_0:\beta=\beta_0$ if $S_n\le c_n$. Construct the *Wald confidence interval* $[p'\hat{\delta}_n-z_{\alpha/2}SE(p'\hat{\delta}_n),\ p'\hat{\delta}_n+z_{\alpha/2}SE(p'\hat{\delta}_n)]$, where $SE(\cdot)$ is the (estimated) standard error of its argument, and z_α is the $1-\alpha$ quantile of the standard normal distribution.

[16] Since β's value is pinned down by the null hypothesis, it is natural to consider inference on the parameters that are estimated under both hypotheses.

The heuristic behind this procedure is as follows. First, we compare S_n to a critical value c_n that tends to 0 slowly. For DGPs with β well separated from its null value's neighborhood, we cannot ensure the asymptotic validity of the Wald confidence interval. In such a case, the pre-test that rejects H_0 with a high probability should prescribe a robust confidence interval, which controls the asymptotic coverage probability regardless of β's value. Since the critical value c_n is shrunk toward 0, we use the Wald confidence interval only if β is in a small local neighborhood of β_0 against which the score test has little power. The shrinkage factor κ_n, therefore, introduces a conservative distortion, which is expected to make the resulting confidence interval's coverage probability above its nominal over a wide range of β values and other features of the DGP.

4 An Empirical Illustration

To show that our inference approach can handle a moderate number of nuisance parameters, we consider an application to the two-player entry game as in Kline and Tamer (2016).

The data come from the second quarter of the 2010 Airline Origin and Destination Survey (DB1B)[17] and contain 7882 markets, formally defined as trips between airports regardless of intermediate stops. There are two types of firms in each market: *LCC* (low cost carriers) and *OA* (other airlines). The binary variable $y_{\ell,i}$ takes value 1 if airline $\ell \in \{\text{LCC, OA}\}$ serves the market i. Airline i's payoff in market i equals

$$y_{\ell,i}(\delta_\ell^{cons} + \delta_\ell^{size} X_{i,size} + \delta_\ell^{pres} X_{\ell,i,pres} + \beta_\ell y_{-\ell,i} + \epsilon_{\ell,i}),$$

where β_ℓ captures the impact of competitor's entry decision, $y_{-\ell,i}$. The empirical question of interest is "whether the LCCs and other airlines compete in a strategic way", which can be considered as a one–sided testing question. The null hypothesis is $\beta_{LCC} = \beta_{OA} = 0$ and the alternative is existence of substitution effects, i.e., $\beta_\ell < 0$.

Regarding observable covariates, in addition to airline–specific intercepts, there are two explanatory variables: *market size* $X_{i,size}$ and *market presence* $X_{\ell,i,pres}$. The market-specific variable $X_{i,size}$ is defined as the population at the endpoints of each trip. The second explanatory variable $X_{\ell,i,pres}$ is airline– and market-specific, and it is defined as the average of the ratios between the number of markets that airline i serves from a given airport and the total number of markets served from that airport by any airline from the two endpoints. It is an excluded variable because presence for airline ℓ only enters ℓ's payoff. In one of our specifications, we treat $X_{m,size}$ and $X_{i,m,pres}$ as continuous variables normalized to the unit interval.[18] In another specification, we let $X_{m,size}$ and $X_{i,m,pres}$ take value 1 if they are above their respective medians and 0 otherwise following the discretization approach in Kline and

[17]The data are available on Brendan Kline's website.

[18]The scale of the two variables without discretization differ significantly, and thus we rescale each variable to be between 0 and 1.

Tamer (2016). The coefficients on the covariates and intercepts are nuisance parameters. In total, we have 6 such parameters.

The results of the analysis are reported in Table 1. When the covariates are discretized, the test rejects the null hypothesis at the 5% level, which is consistent with the finding of Kline and Tamer (2016) whose credible sets for $\beta_\ell, \ell = 1, 2$ do not contain the origin. For comparison, we also consider a specification without discretization since the robust score test can accommodate continuous regressors as well. Under this specification, the result changes drastically; we do not reject the presence of strategic substitution even at the 10% level. This suggests that a model without any strategic interaction effects can potentially explain the observed pattern of market entry once we properly take into account the variation of the continuous covariates.

	p-value
Discretized	0.0002
Not discretized	0.6262

Table 1: p-values of the score test

	$\hat{\delta}^{pres}_{LCC}$	$\hat{\delta}^{size}_{LCC}$	$\hat{\delta}^{cons}_{LCC}$	$\hat{\delta}^{pres}_{OA}$	$\hat{\delta}^{size}_{OA}$	$\hat{\delta}^{cons}_{OA}$
Discretized	1.643	0.795	-2.084	0.388	0.440	0.338
Not discretized	7.102	0.453	-4.111	4.690	1.224	-2.656

Table 2: Estimated values of δ under H_0

Note: $X_{\ell,size}$ and $X_{\ell,i,pres}$ are treated as continuous variables on the unit interval when they are not discretized; $X_{\ell,size}$ and $X_{\ell,i,pres}$ are binary indicators of whether the original variables are above their median or not when they are discretized.

5 Monte Carlo Experiments

5.1 Size and Power of the Score Test

We examine the size and power properties of the score test through simulations. The data generating process is based on Example 1 and is motivated by the empirical illustration in the previous section. There are player-specific covariates $x_i = (x_i^{(1)}, x_i^{(2)})'$, each of which is generated as an independent Rademacher random variable taking values on $\{-1, 1\}$. We then generate $u_i = (u_i^{(1)}, u_i^{(2)})$ from the bivariate standard normal distribution. For each u_i and x_i, we determine the predicted set of outcomes $G(u_i|x_i;\theta)$ based on the payoff functions with $\delta_0 = (\delta_0^{(1)}, \delta_0^{(2)}) = (2, 1.5)'$. We then test

$$H_0 : \beta^{(1)} = \beta^{(2)} = 0, \quad v.s. \quad H_1 : \beta^{(1)} < 0, \beta^{(2)} < 0. \tag{5.1}$$

As discussed earlier, the model is complete under H_0. We estimate δ_0 using the restricted MLE. The sample size is set to 2500, 5000, or 7500. This choice is motivated by the sample size used in the empirical application.

The size of the score test is reported in Table 3. The size of the test is controlled properly with $n = 7500$, but there are small size distortions when the sample size is smaller possibly due to estimation errors associated with components of the estimated information matrix.

Sample size	2500	5000	7500
Size	0.065	0.057	0.048

Table 3: Size of the score test

Under alternative hypotheses, multiple equilibria may be predicted. If this is the case, we select an outcome according to one of the following selection mechanisms. The first design uses a selection mechanism, which selects $(1,0)$ out of $G(u_i|x_i;\theta) = \{(1,0),(0,1)\}$ if an i.i.d. Bernoulli random variable ν_i takes 1. In the second design, we generate data from the least favorable distribution, which draws an independent outcome sequence from the least favorable distribution $Q_{\theta_1} \in \mathcal{Q}_{\theta_1}$.

The power of the score test is calculated against local alternatives with $\beta_1^{(j)} = -h/\sqrt{n}, h > 0$ for $j = 1,2$. For this exercise, we introduce a grid of values for h and generate data described as above. We then compare the rejection frequency of our test to that of the moment-based testing procedure by Bugni et al. (2017). Their test checks if a hypothesized value $(\beta^{(1)}, \beta^{(2)})' = (0,0)'$ is compatible with a set of moment restrictions. Their statistic and bootstrap critical value are calculated using a sample analog of the following moment inequality and equality restrictions

$$
\begin{aligned}
&P(Y = (1,0)|X = x) \geq (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})]\\
&P(Y = (1,0)|X = x) \leq (1 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1\\
&P(Y = (0,0)|X = x) = (1-\Phi_1)(1-\Phi_2)\\
&P(Y = (1,1)|X = x) = \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})\Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}),
\end{aligned}
$$

which are the sharp identifying restrictions that characterize $\mathfrak{q}_\theta$ in (3.4).[19]

Figure 3 shows the rejection frequency of our test under the i.i.d. selection mechanism. It shows that our test outperforms that of the moment-based test by a large margin.

[19] Since the example resembles the specification used in their Monte Carlo experiments, we added minimal changes to their replication code posted on the repository of Quantitative Economics to implement their procedure.

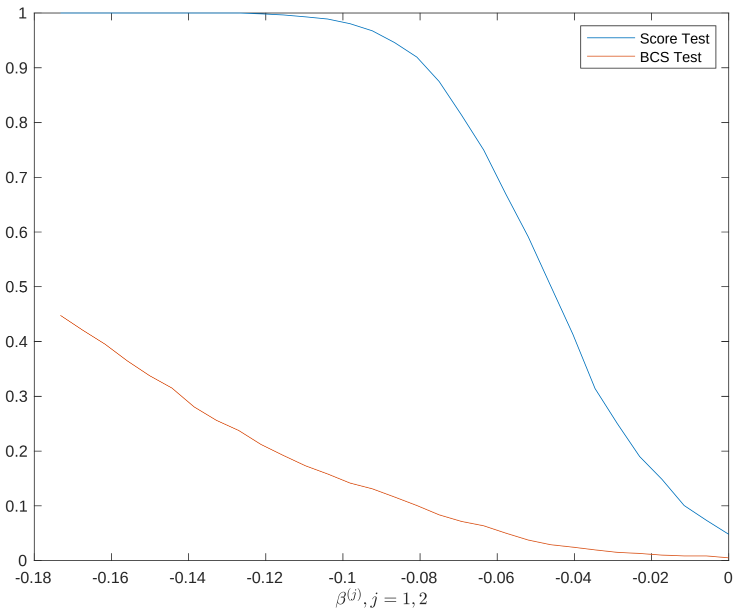


Figure 3: Power of the Score and BCS Tests (Design 1)

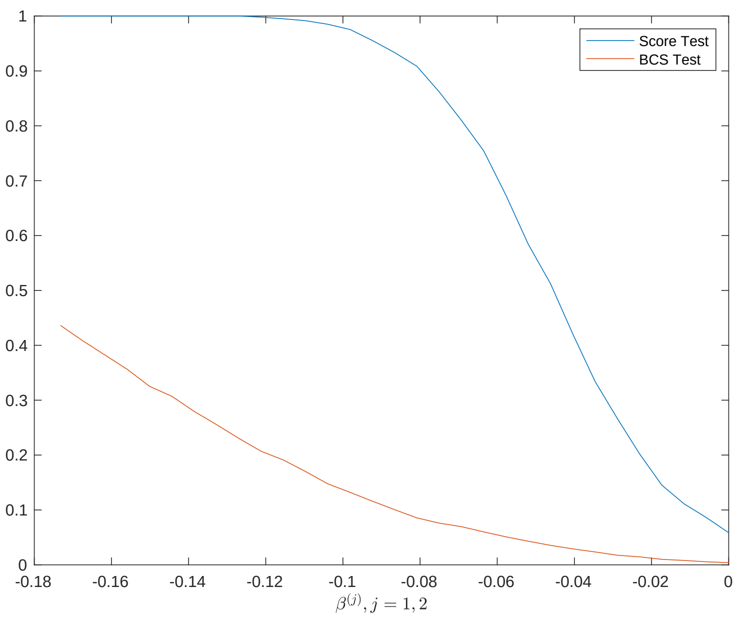


Figure 4: Power of the Score and BCS Tests (Design 2)

6 Concluding remarks

This paper proposes a novel score-based test of model completeness against incompleteness. Our test is attractive in settings where the model involves nuisance parameters, which is common in applications. The score test only requires estimation of nuisance parameters within the restricted model, which is complete. We utilize a point estimator of the nuisance parameters whereby avoiding evaluations of the test statistic over a large number of parameter values. The results of Monte Carlo experiments suggest the score test has an advantage in terms of power over an existing method. An avenue for future research includes a unified theory for the uniform validity of inference for post-model selection procedures that are based on our score test.

References

ALIPRANTIS, C. D. AND K. C. BORDER (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer.

ANDREWS, D. W. K. (1999): "Estimation When a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

ANDREWS, D. W. K. AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

ANDREWS, I., J. ROTH, AND A. PAKES (2019): "Inference for Linear Conditional Moment Inequalities," Discussion Paper, Harvard University.

BARSEGHYAN, L., M. COUGHLIN, F. MOLINARI, AND J. C. TEITELBAUM (2021): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048.

BERA, A. K. AND Y. BILIAS (2001): "Rao's Score, Neyman's $C(\alpha)$ and Silvey's LM tests: An Essay on Historical Developments and Some New Results," *Journal of Statistical Planning and Inference*, 97, 9–44.

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.

BRESNAHAN, T. F. AND P. C. REISS (1991a): "Empirical models of discrete games," *Journal of Econometrics*, 48, 57–81.

——— (1991b): "Entry and Competition in Concentrated Markets," *Journal of Political Economy*, 99, 977–1009.

BUGNI, F., I. CANAY, AND X. SHI (2017): "Inference for Subvectors and Other Functions of Partially Identified Parameters in Moment Inequality Models," *Quantitative Economics*, 8, 1–38.

CANAY, I. A. AND A. M. SHAIKH (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2 of *Econometric Society Monographs*, 271–306.

CHEN, X., T. M. CHRISTENSEN, AND E. TAMER (2018): "Monte Carlo confidence sets for identified sets," *Econometrica*, 86, 1965–2018.

CHESHER, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

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

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

Cox, G. and X. Shi (2020): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," Working Paper.

de Paula, A., S. Richards-Shubik, and E. Tamer (2018): "Identifying Preferences in Networks With Bounded Degree," *Econometrica*, 86, 263–288.

de Paula, Á. and X. Tang (2012): "Inference of Signs of Interaction Effects in Simultaneous Games With Incomplete Information," *Econometrica*, 80, 143–172.

Dempster, A. (1967): "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Mathematical Statistics*, 38, 325–339.

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

Epstein, L., H. Kaido, and K. Seo (2016): "Robust Confidence Regions for Incomplete Models," *Econometrica*, 84, 1799–1838.

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.

Galichon, A. and M. Henry (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, 78, 1264–1298.

Gilboa, I. and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

Haile, P. A. and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111, 1–51.

Heckman, J. J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against The Hypothesis of Spurious State Dependence," *Annales de INSEE*, 227–269.

——— (1987): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence," in *Structural Analysis of Discrete Data With Econometric Applications*, ed. by C. Manski and D. McFadden, MIT Press.

Henry, M., R. Meango, and I. Mourifié (2020): "Revealing Gender-Specific Costs of STEM in an Extended Roy Model of Major Choice," Working Paper.

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

Huber, P. and V. Strassen (1973): "Minimax Tests and Neyman–Pearson Lemma for Capacities," *The Annals of Statistics*, 1, 251–263.

Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 87, 1397–1432.

Kaido, H. and Y. Zhang (2019): "Robust Likelihood-Ratio Tests for Incomplete Economic Models," Working Paper.

Kawai, K. and Y. Watanabe (2013): "Inferring Strategic Voting," *American Economic Review*, 103, 624–62.

Kline, B. and E. Tamer (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366.

Kocherlakota, S. and K. Kocherlakota (1991): "Neyman's $C(\alpha)$ test and Rao's efficient score test for composite hypotheses," *Statistics & Probability Letters*, 491–493.

Lewbel, A. (2007): "Coherency and Completeness of Structural Models Containing a Dummy Endogenous Variable," *International Economic Review*, 48, 1379–1392.

Molinari, F. (2020): "Microeconometrics with Partial Identification," in *Handbook of Econometrics*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7, 355–486.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 4, chap. 36, 2111–2245.

Neyman, J. (1959): "Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics (The Harald Cramér Volume)*, Almquist and Wiksells, Uppsala, Sweden, 213–234.

——— (1979): "C(α) Tests and Their Use," *Sankhyā: The Indian Journal of Statistics, Series A*, 41, 1–21.

Otsu, T., M. Pesendorfer, and Y. Takahashi (2016): "Pooling Data Across Markets in Dynamic Markov Games," *Quantitative Economics*, 7, 523–559.

Pelican, A. and B. S. Graham (2021): "An Optimal Test for Strategic Interaction in Social and Economic Network Formation Between Heterogeneous Agents," Working Paper.

Philippe, F., G. Debs, and J.-Y. Jaffray (1999): "Decision Making with Monotone Lower Probabilities of Infinite Order," *Mathematics of Operations Research*, 24, 767–784.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002.

Shafer, G. (1976): *A Mathematical Theory of Evidence*, Princeton University Press.

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955.

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

Silvapulle, M. J. and P. Silvapulle (1995): "A Score Test Against One-Sided Alternatives," *Journal of the American Statistical Association*, 90, 342–349.

Talagrand, M. (1994): "Sharper Bounds for Gaussian and Empirical Processes," *The Annals of Probability*, 22, 28–76.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165.

Train, K. (2009): *Discrete Choice Methods with Simulation*, Cambridge University Press, 2 ed.

van der Vaart, A. (2000): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics.*, Springer, New York.

Wasserman, L. A. (1990): "Belief Functions and Statistical Inference," *Canadian Journal of Statistics*, 18, 183–196.

Wollmann, T. G. (2018): "Trucks Without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," *American Economic Review*, 108, 1364–1406.

Wooldridge, J. M. (2005): "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.

——— (2014): "Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 182, 226–234.

——— (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445.

A Notation and Preliminaries

For ease of reference, the following list includes notation and definitions that will be used throughout the Appendices:

$a \lesssim b$	$a \leq Mb$ for some constant M.
$\|\cdot\|_{op}$	the operator norm for linear mappings.
$\|\cdot\|_{\mathcal{F}}$	the supremum norm over $\mathcal{F}$.
$N(\epsilon, \mathcal{F}, \|\cdot\|)$	covering number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$.
$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$	bracketing number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$.
$X_n \overset{P^n}{\rightsquigarrow} X$	X_n weakly converges to X under $\{P^n\}$

Table 4: Notation and definitions

Let Ω be a compact metric space and let Σ_Ω denote its Borel σ-algebra. Let $\mathcal{K}(\Omega)$ be the set of compact subsets of Ω endowed with the Hausdorff metric. Let $\mathcal{C}(\Omega)$ be the set of continuous functions on Ω. Let $\Delta(\Omega)$ be the set of Borel probability measures on Ω endowed with the weak topology.

A set function ν^* is said to be a *capacity* if ν^* satisfies the following conditions:

(i) $\nu^*(\varnothing) = 0, \nu^*(\Omega) = 1,$

(ii) $A \subset B \Rightarrow \nu^*(A) \leq \nu^*(B),$ for all $A, B \in \Sigma_\Omega$.

(iii) $A_n \uparrow A \Rightarrow \nu^*(A_n) \uparrow \nu^*(A)$, for all $\{A_n, n \geq 1\} \subset \Sigma_\Omega$ and $A \in \Sigma_\Omega$.

(iv) $F_n \downarrow F, F_n$ closed $\Rightarrow \nu^*(F_n) \downarrow \nu^*(F)$.

One may define integral operations with respect to capacities as follows. Let $f : \Omega \to \mathbb{R}$ be a measurable function. The *Choquet integral* of f with respect to ν is defined by

$$\int f d\nu \equiv \int_{-\infty}^{0} (\nu(\{\omega : f(\omega) \geq t\}) - \nu(\Omega))dt + \int_{0}^{\infty} \nu(\{\omega : f(\omega) \geq t\})dt, \tag{A.1}$$

where the integrals on the right hand side are Riemann integrals. A capacity ν is said to be *monotone of order* k or, for short, k*-monotone* if for any $A_i \subset S, i = 1 \cdots, k$,

$$\nu\big(\cup_{i=1}^{k} A_i\big) \geq \sum_{I \subseteq \{1,\cdots,k\}, I \neq \varnothing} (-1)^{|I|+1} \nu\big(\cap_{i\in I} A_i\big). \tag{A.2}$$

Conjugate $\nu^*(A) = 1 - \nu(A^c)$ is then called a k*-alternating* capacity. A capacity that satisfies (A.2) is called an *infinitely monotone capacity* or a *belief function*. Capacities are used in various areas of statistics (Dempster, 1967; Shafer, 1976; Wasserman, 1990) and economics (Gilboa and Schmeidler, 1989).

The following result, known as Choquet's theorem, states that a random closed set K following a distribution M induces a belief function, and it follows from Theorems 1-3 in Philippe et al. (1999).

Lemma A.1. *Let Ω be a Polish space. Let M be a probability measure on $\mathcal{K}(\Omega)$. Let $\mathcal{P} = \{P \in \Delta(\Omega) : P = \int P_K dM(K), P_K \in \Delta(K)\}$. Then, $\nu(\cdot) = \inf_{P\in\mathcal{P}} P(\cdot)$ is a belief function and satisfies*

$$\nu(A) = M(\{K \subset A\}). \tag{A.3}$$

In our setting, we apply the lemma above with a random subset of $\mathcal{Y} \times \mathcal{X}$. Namely, we take $K = G(u|X;\theta) \times \{X\}$, and M is the law of K induced by u's conditional distribution F_θ and X's marginal distribution q_x. We then denote the induced belief function by ν_θ and its conjugate ν^*_θ (see (C.6)-(C.7) below).

B Details on the Examples

B.1 Discrete Games of Complete Information

We focus on Example 1 below, but the analysis of Example 2 is similar.

Model restrictions and Assumption 1

The upper and lower probabilities of all singleton events are tabulated in Table 5. In this example, they constitute the sharp identifying restrictions (Galichon and Henry, 2011).

Table 5: Upper and Lower Probability Bounds in Game with Nuisance Parameters

Event A	$\nu_\theta(A) = \min P(A)$	$\nu^*_\theta(A) = \max P(A)$
$\{(0,0)\}$	$(1-\Phi_1)(1-\Phi_2)$	$(1-\Phi_1)(1-\Phi_2)$
$\{(1,1)\}$	$\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})$	$\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})$
$\{(1,0)\}$	$(1-\Phi_2)\Phi_1+\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]$	$(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))\Phi_1$
$\{(0,1)\}$	$(1-\Phi_1)\Phi_2+\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]$	$(1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)}))\Phi_2$

As argued in Section 3.2, the model's prediction reduces to (3.16) when $\beta^{(1)} = \beta^{(2)} = 0$, which implies a unique density in (3.14). Therefore, Assumption 1 (i) holds. For Assumption 1 (ii), it suffices to show that $\mathcal{Q}_{\theta_0}$ and $\mathcal{Q}_{\theta_1}$ are disjoint. For this, consider the event $\{(1,1)\}$. Table 5 suggests

$$\nu_{\theta_0}(\{(1,1)\}|x) = \Phi(x^{(1)\prime}\delta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}), \tag{B.1}$$

whereas

$$\nu^*_{\theta_1}(\{(1,1)\}|x) = \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}). \tag{B.2}$$

This means $\nu^*_{\theta_1}(\{(1,1)\}|x) < \nu_{\theta_0}(\{(1,1)\}|x)$ whenever $\beta^{(j)} < 0, j = 1,2$. Hence, $\mathcal{Q}_{\theta_0}$ and $\mathcal{Q}_{\theta_1}$ are disjoint.

Computing the LFP

The LF density q_θ is given by

$$q_\theta(0,0|x) = (1-\Phi_1)(1-\Phi_2) \tag{B.3}$$

$$q_\theta(1,1|x) = \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}) \tag{B.4}$$

$$q_\theta(1,0|x) = \begin{cases} \Phi_1(1-\Phi_2) + \frac{\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{\Phi_1+\Phi_2-2\Phi_1\Phi_2} & \theta\in\Theta_1(x) \\ \Phi_1(1-\Phi_2)+\Phi(x^{(1)\prime}\delta_1+\beta^{(1)})[\Phi_2-\Phi(x^{(2)\prime}\delta_2+\beta^{(2)})] & \theta\in\Theta_2(x) \\ \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})) & \theta\in\Theta_3(x) \end{cases}, \tag{B.5}$$

where $\Theta_j(x), j = 1,2,3$ are given by

$$\Theta_1(x) = \{\theta : \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})) \geq \frac{z_1z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2} \tag{B.6}$$

$$\frac{z_1z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2} \geq \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]\}$$

$$\Theta_2(x) = \{\theta : \frac{z_1z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2} < \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]\} \tag{B.7}$$

$$\Theta_3(x) = \{\theta : \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})) < \frac{z_1z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}\}, \tag{B.8}$$

where

$$z_1 = \Phi_1(1-\Phi_2)$$
$$z_2 = \Phi_2(1-\Phi_1) + \Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}).$$

Below, we outline how to obtain this density from the convex program in (3.6)-(3.8).

As discussed in the text, q_{θ_0} is determined by the four equality restrictions (3.10)-(3.13). Therefore, it remains to solve the convex program in (3.6)-(3.8) for q_1. For this, we can reduce the number of control variables. First, Table 5 implies

$$q_{\theta_1}(0,0|x) = (1-\Phi_1)(1-\Phi_2) \tag{B.9}$$
$$q_{\theta_1}(1,1|x) = \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}). \tag{B.10}$$

Hence, the remaining free components of q_1 are $q_1(1,0|x)$ and $q_1(0,1|x)$. Let $\omega = q_1(1,0|x)$. We may then express the other component as

$$q_1(0,1|x) = 1 - q_1(0,0|x) - q_1(1,1|x) - \omega = \Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}) - \omega.$$

Hence, to solve(3.6)-(3.8), it suffices to choose $\omega = q_1(1,0|x)$ optimally in the following problem:

$$\min_{\omega\in[0,1]} -\ln\Big(\frac{z_1}{z_1+\omega}\Big)(z_1+\omega) - \ln\Big(\frac{(1-\Phi_1)\Phi_2}{z_2-\omega}\Big)(z_2-\omega) \tag{B.11}$$

$$s.t. \omega - (1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))\Phi_1 \leq 0 \tag{B.12}$$

$$(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})] - \omega \leq 0. \tag{B.13}$$

Let the Lagrangian be

$$\mathcal{L}(\omega,\lambda) = -\ln\Big(\frac{z_1}{z_1+\omega}\Big)(z_1+\omega) - \ln\Big(\frac{(1-\Phi_1)\Phi_2}{z_2-\omega}\Big)(z_2-\omega) - \lambda_1((1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))\Phi_1 - \omega)$$
$$- \lambda_2(\omega - (1-\Phi_2)\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]).$$

The Karush-Kuhn-Tucker (KKT) conditions are

$$-\ln\left(\frac{z_1}{z_1+\omega}\right) + \ln\left(\frac{\Phi_2(1-\Phi_1)}{z_2-\omega}\right) + \lambda_1 - \lambda_2 = 0 \tag{B.14}$$

$$\lambda_1\Big(\Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})) - \omega\Big) \geq 0 \tag{B.15}$$

$$\lambda_2\Big(\omega - (1-\Phi_2)\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]\Big) \geq 0 \tag{B.16}$$

$$\lambda_1, \lambda_2 \geq 0. \tag{B.17}$$

Below, we consider three subcases depending on the value of the Lagrange multipliers.

Case 1 ($\lambda_1 = \lambda_2 = 0$) The FOC in (B.14) with $\lambda_1 = \lambda_2 = 0$ identifies the solution $q_{\theta_1}(1,0|x)$ as follows:

$$\omega = q_{\theta_1}(1,0|x) = \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}$$
$$= \frac{\Phi_1(1-\Phi_2)[\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \quad \text{(B.18)}$$

This implies

$$q_{\theta_1}(0,1|x) = \Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}) - \omega \quad \text{(B.19)}$$
$$= \frac{\Phi_2(1-\Phi_1)[\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \quad \text{(B.20)}$$

Substituting the value of ω into its bounds, we obtain the following restrictions:

$$\Phi_1(1 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})) - \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2} \geq 0 \quad \text{(B.21)}$$
$$\frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2} - \Phi_1(1-\Phi_2) - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})] \geq 0. \quad \text{(B.22)}$$

We let $\Theta_1(x)$ denote the set of parameter values that satisfy (B.21)-(B.22).

Case 2 ($\lambda_1 = 0$, $\lambda_2 > 0$). By $\lambda_2 > 0$ and (B.16), we obtain

$$\omega = q_{\theta_1}(1,0|x) = (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})],$$

and $q_{\theta_1}(0,1|x) = (1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)}))\Phi_2$. Note that $\lambda_2 > 0$ iff

$$\frac{z_1}{z_1 + \omega} < \frac{\Phi_2(1-\Phi_1)}{z_2 - \omega},$$

which is equivalent to

$$(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})] > \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \quad \text{(B.23)}$$

We let $\Theta_2(x)$ denote the set of parameter values that satisfy (B.23).

Case 3 ($\lambda_1 > 0$, $\lambda_2 = 0$). By $\lambda_1 > 0$ and (B.15), we obtain

$$\omega = q_{\theta_1}(1,0|x) = (1 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1,$$

and hence $q_{\theta_1}(0,1|x) = (1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})]$. Note that $\lambda_1 > 0$ iff

$$\frac{z_1}{z_1 + \omega} > \frac{\Phi_2(1-\Phi_1)}{z_2 - \omega},$$

which is equivalent to

$$(1 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1 < \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \quad \text{(B.24)}$$

We let $\Theta_3(x)$ denote the set of parameter values that satisfy (B.24).

Score:

We let $s_\theta = (s_{\beta^{(1)}}, s_{\beta^{(2)}}, s_{\delta^{(1)}}, s_{\delta^{(2)}})'$. Each component will be of the form

$$s_\vartheta(y|x) = \sum_{\bar{y}\in\mathcal{Y}} 1\{y = \bar{y}\} z_\vartheta(\bar{y}|x),\ \ \vartheta \in \{\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}\}, \tag{B.25}$$

where $z_\vartheta(\bar{y}|x)$ is the partial derivative of $\ln p_\theta(\bar{y}|x)$ with respect to ϑ, which is well-defined if θ is in $\Theta_2(x)$, $\Theta_3(x)$, or in the interior of $\Theta_1(x)$. Let

$$r_h(y,x) \equiv (\sqrt{q_{\theta+h}(y|x)} - \sqrt{q_\theta(y|x)} - \frac{1}{2}h' s_\theta(y|x)\sqrt{q_\theta(y|x)})^2. \tag{B.26}$$

Suppose $\theta \in \Theta_2(x)$. By (B.7), $\theta + h \in \Theta_2(x)$ for $\|h\|$ small enough. Then, pointwise, $r_h(y,x) = o(\|h^2\|)$ because $s_\theta(y|x) = 2\frac{1}{\sqrt{q_\theta(y|x)}}\frac{\partial}{\partial\theta}\sqrt{q_\theta(y|x)} = \frac{\partial}{\partial\theta}\ln q_\theta(y|x)$. The same argument applies when $\theta \in \Theta_3(x)$ or $\theta \in \mathrm{int}(\Theta_1(x))$. The only case this argument does not apply is when θ is on the boundary between $\Theta_2(x)$ and $\Theta_1(x)$ (or between $\Theta_3(x)$ and $\Theta_1(x)$). For example, suppose θ is on the boundary between $\Theta_2(x)$ and $\Theta_1(x)$. Then, we may have $\theta + h \in \Theta_2(x)$ for all h with $\|h\| > 0$ but $\theta \in \Theta_1$. Then, the pointwise argument above does not apply. However, θ being on the boundary between the two sets means

$$\frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2} = \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})].$$

If x contains a continuous component (e.g. distance from headquarters/distribution center), the set of x's satisfying above has measure 0, $r_h(y,x)$ is bounded on the set, and hence it does not affect the integral in Assumption 2. Hence, Assumption 2 holds.

For completeness, the functional form of $z_\vartheta(\bar{y}|x)$ is derived below for each $\bar{y} \in \mathcal{Y}$ and $\vartheta \in \{\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}\}$. Across all subcases analyzed in the previous section, the form of $q_\theta(0,0|x)$ and $q_\theta(1,1|x)$ remains the same. We calculate score functions first by taking the pointwise derivative of $\ln q_\theta(0,0|x)$ and $\ln q_\theta(1,1|x)$. This yields

$$z_{\beta^{(1)}}(0,0|x) = 0, \quad z_{\beta^{(2)}}(0,0|x) = 0, \quad z_{\delta^{(1)}}(0,0|x) = -\phi_1 x^{(1)\prime}/(1-\Phi_1), \quad z_{\delta^{(2)}}(0,0|x) = -\phi_2 x^{(2)\prime}/(1-\Phi_2)$$

$$z_{\beta^{(1)}}(1,1|x) = \frac{\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\beta^{(2)}}(1,1|x) = \frac{\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}$$

$$z_{\delta^{(1)}}(1,1|x) = \frac{\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})x^{(1)}}{\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\delta^{(2)}}(1,1|x) = \frac{\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})x^{(2)}}{\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})},$$

where $\phi_j = \phi(x^{(j)\prime}\delta^{(j)})$, $j = 1, 2$.

Next, we derive $z_\vartheta(1,0|x)$ and $z_\vartheta(0,1|x)$.

Case 1: Suppose $\theta \in \Theta_1(x)$. By taking the pointwise derivative of $\ln q_\theta$ in (B.5), one can obtain

$$z_{\delta^{(1)}}(1,0|x) = \frac{\phi_1 x^{(1)}}{\Phi_1} + \frac{\phi_1 x^{(1)} - \phi_1\Phi_2 x^{(1)} - \phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})x^{(1)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_1 x^{(1)}(1-2\Phi_2)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\delta^{(2)}}(1,0|x) = \frac{-\phi_2 x^{(2)}}{1-\Phi_2} + \frac{\phi_2 x^{(2)} - \phi_2\Phi_1 x^{(2)} - \phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})x^{(2)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_2 x^{(2)}(1-2\Phi_1)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\beta^{(1)}}(1,0|x) = \frac{-\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}$$

$$z_{\beta^{(2)}}(1,0|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}.$$

Similarly,

$$z_{\delta^{(1)}}(0,1|x) = \frac{-\phi_1 x^{(1)}}{1-\Phi_1} + \frac{\phi_1 x^{(1)} - \phi_1\Phi_2 x^{(1)} - \phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})x^{(1)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_1 x^{(1)}(1-2\Phi_2)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\delta^{(2)}}(0,1|x) = \frac{\phi_2 x^{(2)}}{\Phi_2} + \frac{\phi_2 x^{(2)} - \phi_2\Phi_1 x^{(2)} - \phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})x^{(2)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_2 x^{(2)}(1-2\Phi_1)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\beta^{(1)}}(0,1|x) = \frac{-\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}$$

$$z_{\beta^{(2)}}(0,1|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}.$$

Case 2: Suppose $\theta \in \Theta_2(x)$. Similarly to the analysis in Case 1, we may obtain

$$z_{\delta^{(1)}}(1,0|x) = \frac{x^{(1)}(1-\Phi_2)\phi_1 + x^{(1)}\Phi_2\phi(x^{(1)\prime}\delta^{(1)} + \beta_1) - x^{(1)}\Phi(x^{(2)\prime}\delta^{(2)} + \beta_2)\phi(x^{(1)\prime}\delta^{(1)} + \beta_1)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\delta^{(2)}}(1,0|x) = \frac{-x^{(2)}\Phi_1\phi_2 + x^{(2)}\phi_2\Phi(x^{(1)\prime}\delta^{(1)} + \beta_1) - x^{(2)}\Phi(x^{(1)\prime}\delta^{(1)} + \beta_1)\phi(x^{(2)\prime}\delta^{(2)} + \beta_2)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\beta^{(1)}}(1,0|x) = \frac{(\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}))\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\beta^{(2)}}(1,0|x) = \frac{\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]},$$

and

$$z_{\delta^{(1)}}(0,1|x) = -\frac{x^{(1)}\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{1-\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\delta^{(2)}}(0,1|x) = \frac{x^{(2)}\phi_2}{\Phi_2}$$

$$z_{\beta^{(1)}}(0,1|x) = -\frac{\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{1-\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\beta^{(2)}}(0,1|x) = 0.$$

Case 3: Suppose $\theta \in \Theta_3(x)$. Similarly to the previous two cases, we may obtain

$$z_{\delta^{(1)}}(1,0|x) = x^{(1)}\phi_1/\Phi_1, \quad z_{\delta^{(2)}}(1,0|x) = \frac{-x^{(2)}\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}$$

$$z_{\beta^{(1)}}(1,0|x) = 0, \quad z_{\beta^{(2)}}(1,0|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})},$$

and

$$z_{\delta^{(1)}}(0,1|x) = \frac{-x^{(1)}\Phi_2\phi_1 + \Phi(x^{(2)\prime}+\delta^{(2)})x^{(1)}(\phi_1 - \phi(x^{(1)\prime}\delta^{(1)}+\delta^{(1)}))}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]}$$

$$z_{\delta^{(2)}}(0,1|x) = \frac{x^{(2)}(1-\Phi_1)\phi_2 + x^{(2)}(\Phi_1 - \Phi(x^{(1)\prime}\delta_1+\beta_1))\phi(x^{(2)\prime}\delta_2+\beta_2)}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]}$$

$$z_{\beta^{(1)}}(0,1|x) = \frac{-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})\phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]}$$

$$z_{\beta^{(2)}}(0,1|x) = \frac{(\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)}))\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]}.$$

B.2 Triangular Model with an Incomplete Control Function

We derive an explicit form of G below. Suppose $d_i = 1$ first. By (2.5)-(2.6), $y_i = 0$ if

$$u_i < -\alpha - w_i'\eta - \beta v_i, \quad \text{for some } v_i \in [-z_i'\gamma, \infty). \tag{B.27}$$

Suppose $\beta > 0$. Then, this event is equivalent to $u_i \in (-\infty, -\alpha - w_i'\eta + \beta z_i'\gamma)$.

On the other hand, $y_i = 1$ if

$$u_i \geq -\alpha - w_i'\eta - \beta v_i, \quad \text{for some } v_i \in [-z_i'\gamma, \infty). \tag{B.28}$$

Again assuming $\beta > 0$, this means that $y_i = 0$ is consistent with the model whenever $u_i > -\infty$. Let $x_i = (d_i, w_i, z_i)$. These predictions can be summarized as

$$G(u_i|1, w_i, z_i; \theta) = \begin{cases} \{1\} & u_i \geq -\alpha - w_i'\eta + \beta z_i'\gamma \\ \{0,1\} & u_i < -\alpha - w_i'\eta + \beta z_i'\gamma. \end{cases} \tag{B.29}$$

Now suppose $d_i = 0$ implying $v_i \in (-\infty, -z_i'\gamma)$. Repeating a similar analysis yields the following correspondence

$$G(u_i|0, w_i, z_i; \theta) = \begin{cases} \{0\} & u_i \leq -w_i'\eta + \beta z_i'\gamma \\ \{0,1\} & u_i > -w_i'\eta + \beta z_i'\gamma. \end{cases} \tag{B.30}$$

These predictions are summarized in Figure 5.

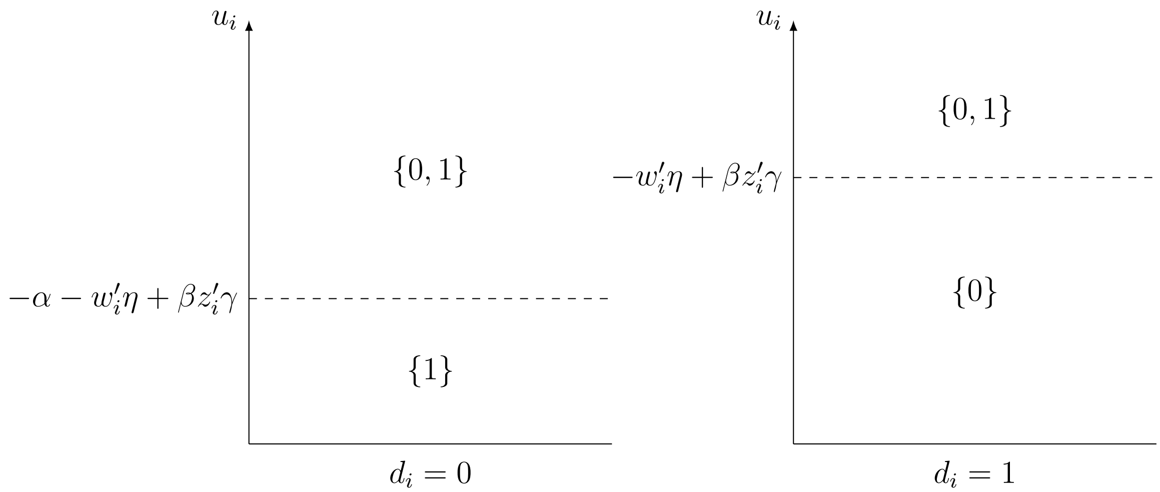


Figure 5: The set of predicted outcomes $G(u|x;\theta)$ of the model when $\beta > 0$

Model restrictions and Assumption 1

Assumption 1 (i) holds because, as argued in Section 3.2, the model's prediction under the null hypothesis becomes

$$g(u_i|d_i, w_i, z_i; \theta) = 1\{\alpha d_i + w_i'\eta + u_i \geq 0\}, \tag{B.31}$$

which induces a unique conditional density for y_i. Suppose $u_i \sim N(0,1)$.[20] Assumption 1 (ii) holds as long as z_i's support is rich enough so that $z_i'\gamma < 0$ with positive probability. For this, we demonstrate that there exists an event A such that $\nu_{\theta_1}(A|x) > \nu^*_{\theta_0}(A|x)$ for some value of $x = (d, w, z)$. For this, take $A = \{1\}$ and suppose $d_i = 0$. Under the null hypothesis, the conditional probability of $y_i = 1$ is uniquely determined as $\nu^*_{\theta_0}(1|d_i = 0, w_i, z_i) = \Phi(w_i'\eta)$. When $\beta > 0$, (B.29) implies

$$\nu_{\theta_1}(\{1\}|d_i = 0, w_i, z_i) = \Phi(w_i'\eta - \beta z_i'\gamma), \tag{B.32}$$

which is greater than $\nu^*_{\theta_0}(1|d_i = 0, w_i, z_i)$ for values of z_i such that $z_i'\gamma < 0$. Hence, $\mathfrak{q}_{\theta_0}$ and $\mathfrak{q}_{\theta_1}$ are disjoint.

Computing LFP and score

For any $\theta = (\beta, \delta)$ with $\beta > 0$, the set $\mathfrak{q}_\theta$ of densities compatible with θ is then characterized by the following inequalities

$$q(0|d = 0, w, z) \geq \Phi(-w_i'\eta + \beta z_i'\gamma) \tag{B.33}$$
$$q(1|d = 0, w, z) \geq 0, \tag{B.34}$$

and

$$q(0|d = 1, w, z) \geq 0 \tag{B.35}$$
$$q(1|d = 1, w, z) \geq 1 - \Phi(-\alpha - w_i'\eta + \beta z_i'\gamma). \tag{B.36}$$

[20] Here, we normalize the scale by setting the variance of u_i to 1. Other choices of normalization are also possible.

Suppose $d = 0$. Let $z = q_1(0|d = 0, w, z)$. Then, the convex program in (3.6)-(3.8) can be written as

$$\min_{(q_0,q_1)} \ln\Big(\frac{q_0(0|x) + z}{q_0(0|x)}\Big)(q_0(0|x) + z) + \Big(\frac{1 - q_0(0|x) + 1 - z}{1 - q_0(0|x)}\Big)(q_0(1|x) + 1 - z) \tag{B.37}$$

$$s.t.\ q_0(y|x) = 1 - \Phi(w'\eta) \tag{B.38}$$

$$z \geq \Phi(-w'\eta + \beta z'\gamma) \tag{B.39}$$

$$1 - z \geq 0, \tag{B.40}$$

where (B.38) is due to the completeness of the model under the null hypothesis (see (3.31)) and $d = 0$. Note that (B.40) is redundant since q_1 being in the probability simplex is implicitly assumed. The KKT conditions associated with the program is therefore

$$\ln \frac{q_0(0|x) + z}{q_0(0|x)} - \ln \frac{2 - q_0(0|x) - z}{1 - q_0(0|x)} - \lambda = 0 \tag{B.41}$$

$$\lambda(\Phi(-w'\eta + \beta z'\gamma) - z) \tag{B.42}$$

$$\lambda \geq 0 \tag{B.43}$$

where $q_0(y|x) = 1 - \Phi(w'\eta)$. There are two cases to consider.

Case 1 ($\lambda = 0$): When $\lambda = 0$, (B.41) implies $z = q_0(y|x) = 1 - \Phi(w'\eta)$. This holds when $z = 1 - \Phi(w'\eta) \geq \Phi(-w'\eta + \beta z'\gamma)$.

Case 2 ($\lambda > 0$): When $\lambda > 0$, $z = \Phi(-w'\eta + \beta z'\gamma)$ by (B.39). This occurs when

$$\lambda = \ln \frac{q_0(0|x) + z}{q_0(0|x)} - \ln \frac{2 - q_0(0|x) - z}{1 - q_0(0|x)} > 0, \tag{B.44}$$

which is equivalent to

$$\Phi(-w'\eta + \beta z'\gamma) > q_0(y|x) = 1 - \Phi(w'\eta). \tag{B.45}$$

In sum, we have

$$q_{\theta_1}(0|d = 0, w, z) = \begin{cases} 1 - \Phi(w'\eta) & \text{if } \Phi(-w'\eta + \beta z'\gamma) \leq 1 - \Phi(w'\eta), \\ \Phi(-w'\eta + \beta z'\gamma) & \text{if } \Phi(-w'\eta + \beta z'\gamma) > 1 - \Phi(w'\eta), \end{cases} \tag{B.46}$$

and $q_{\theta_1}(1|d = 0, w, z) = 1 - q_{\theta_1}(0|d = 0, w, z)$.

Repeating a similar analysis for $d = 1$ yields

$$q_{\theta_1}(1|d = 1, w, z) = \begin{cases} \Phi(\alpha + w'\eta) & \text{if } 1 - \Phi(-\alpha - w'\eta + \beta z'\gamma) \leq \Phi(\alpha + w'\eta), \\ 1 - \Phi(-\alpha - w'\eta + \beta z'\gamma) & \text{if } 1 - \Phi(-\alpha - w'\eta + \beta z'\gamma) > \Phi(\alpha + w'\eta), \end{cases} \tag{B.47}$$

and $q_{\theta_1}(0|d = 1, w, z) = 1 - q_{\theta_1}(1|d = 1, w, z)$.

Recalling $\beta > 0$ and Φ is strictly increasing, we may summarize (B.46)-(B.47) as follows

$$q_{\theta_1}(0|d = 0, w, z) = \begin{cases} \Phi(-w'\eta) & \text{if } z'\gamma \leq 0, \\ \Phi(-w'\eta + \beta z'\gamma) & \text{if } z'\gamma > 0, \end{cases} \tag{B.48}$$

$$q_{\theta_1}(1|d = 0, w, z) = \begin{cases} 1 - \Phi(-w'\eta) & \text{if } z'\gamma \leq 0, \\ 1 - \Phi(-w'\eta + \beta z'\gamma) & \text{if } z'\gamma > 0, \end{cases} \tag{B.49}$$

and

$$q_{\theta_1}(0|d=1,w,z) = \begin{cases} \Phi(-\alpha - w'\eta) & \text{if } z'\gamma \geq 0, \\ \Phi(-\alpha - w'\eta + \beta z'\gamma) & \text{if } z'\gamma < 0, \end{cases} \tag{B.50}$$

$$q_{\theta_1}(1|d=1,w,z) = \begin{cases} 1 - \Phi(-\alpha - w'\eta) & \text{if } z'\gamma \geq 0, \\ 1 - \Phi(-\alpha - w'\eta + \beta z'\gamma) & \text{if } z'\gamma < 0. \end{cases} \tag{B.51}$$

The corresponding score function with respect to β is

$$s_\beta(0|d=0,w,z) = \begin{cases} 0 & \text{if } z'\gamma \leq 0, \\ \frac{\phi(-w'\eta+\beta z'\gamma)}{\Phi(-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma > 0, \end{cases} \tag{B.52}$$

$$s_\beta(1|d=0,w,z) = \begin{cases} 0 & \text{if } z'\gamma \leq 0, \\ -\frac{\phi(-w'\eta+\beta z'\gamma)}{\Phi(-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma > 0, \end{cases} \tag{B.53}$$

and

$$s_\beta(0|d=1,w,z) = \begin{cases} 0 & \text{if } z'\gamma \geq 0, \\ \frac{\phi(-\alpha-w'\eta+\beta z'\gamma)}{1-\Phi(-\alpha-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma < 0, \end{cases} \tag{B.54}$$

$$s_\beta(1|d=1,w,z) = \begin{cases} 0 & \text{if } z'\gamma \geq 0, \\ -\frac{\phi(-\alpha-w'\eta+\beta z'\gamma)}{1-\Phi(-\alpha-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma < 0. \end{cases} \tag{B.55}$$

B.3 Panel Dynamic Discrete Choice Models

For each t, let $u_{it} = \alpha_i + \epsilon_{it}$. We explicitly derive a form of G below. Note that, $y_i = (y_{i1}, y_{i2}) = (0,0)$ occurs if

$$u_{i1} < -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda, \tag{B.56}$$

which follows from (2.10)-(2.11) or

$$u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda, \tag{B.57}$$

which follows from (2.12)-(2.13). When $\beta \geq 0$, the union of the two events reduces to (B.56).

Similarly, $y = (0,1)$ occurs if

$$u_{i1} < -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda, \tag{B.58}$$

or

$$u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda. \tag{B.59}$$

When $\beta \geq 0$, the union of the two events reduces (B.58).

The outcome $y = (1,0)$ occurs if

$$u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda - \beta, \tag{B.60}$$

or

$$u_{i1} \geq -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda - \beta, \tag{B.61}$$

When $\beta \geq 0$, the union of the two events reduces to (B.61).

The outcome $y = (1,1)$ occurs if

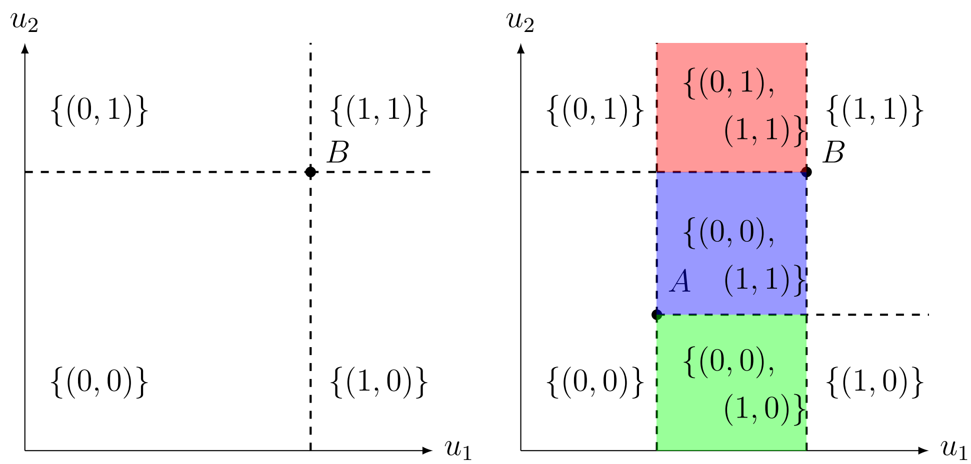
$$u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \tag{B.62}$$

or

$$u_{i1} \geq -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \tag{B.63}$$

When $\beta \geq 0$, the union of the two events reduces to (B.63). These predictions are summarized in Figure 6.

Figure 6: Level sets of $u \mapsto G(u|x;\theta)$ when $\beta \geq 0$



Note: The level sets of G when $\beta = 0$ (left) and $\beta > 0$ (right). $A = (-x_{i1}'\lambda - \beta, -x_{i2}'\lambda - \beta)$; $B = (-x_{i1}'\lambda, -x_{i2}'\lambda)$.
Multiple outcome values are predicted in each of the red, blue, and green regions.

The correspondence can therefore be written as

$$G(u_i|x_i;\theta) = \begin{cases} \{(0,0)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda, \\ \{(0,1)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda, \\ \{(1,0)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda - \beta, \\ \{(1,1)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,0)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \leq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ -x_{i2}'\lambda - \beta \leq u_{i2} < -x_{i2}'\lambda, \\ \{(0,1),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda. \end{cases} \tag{B.64}$$

A similar analysis can be done for the setting with $\beta \leq 0$, which we omit for brevity.

Model restrictions and Assumption 1

Assumption 1 (i) holds because, as argued in (3.18), the model makes a complete prediction with the following reduced-form function when $\beta = 0$:

$$g(u_i|x_i;\theta) = \begin{bmatrix} 1\{x_{i1}'\lambda + \alpha_i + \epsilon_{i1} \geq 0\} \\ 1\{x_{i2}'\lambda + \alpha_i + \epsilon_{i2} \geq 0\} \end{bmatrix}. \tag{B.65}$$

Assumption 1 (ii) holds if u follows a distribution F that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. We show below, when $\beta > 0$, there exists an event A such that $\nu_{\theta_1}(A) > \nu^*_{\theta_0}(A)$ for all $\theta_1 \in \Theta_1$. For example, take $A = \{(1,1)\}$. As shown on the left panel of Figure 6, the probability of $\{(1,1)\}$ is uniquely determined when $\beta = 0$. Therefore, the upper bound on the probability of $\{(1,1)\}$ is

$$\nu^*_{\theta_0}(\{(1,1)\}|x) = F(u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda). \tag{B.66}$$

When $\beta > 0$, the lower bound on the probability of the same event is

$$\nu_{\theta_1}(\{(1,1)\}|x) = F(u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda - \beta), \tag{B.67}$$

which exceeds $\nu^*_{\theta_0}(\{(1,1)\}|x)$ as long as F is absolutely continuous. This means $\mathfrak{q}_{\theta_1}$ and $\mathfrak{q}_{\theta_0}$ are disjoint.

The analysis of the LFP and score is similar to that of discrete games. For brevity, we omit details.

C Lemmas and Proofs

This section is organized as follows. In Section C.1, we show $\sqrt{n}$-consistency of $\hat{\delta}_n$ by extending standard arguments for extremum estimators to locally incomplete models. In Section C.2, we use the results in C.1 to show results on the asymptotic size of our score test.

C.1 $\sqrt{n}$-consistency of $\hat{\delta}_n$

Lemma C.1. *Suppose Assumption 4 holds. Then for any bounded function $g : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$,*

$$\Big|\int g d\nu^*_\theta - \int g d\nu^*_{\theta'}\Big| \leq C'\|\theta - \theta'\|,\ \forall \theta, \theta' \in \Theta, \tag{C.1}$$

and

$$\Big|\int g d\nu_\theta - \int g d\nu_{\theta'}\Big| \leq C'\|\theta - \theta'\|,\ \forall \theta, \theta' \in \Theta, \tag{C.2}$$

for some $C' > 0$.

Proof. Note that

$$\begin{aligned}\int g d\nu^*_\theta &= \int \max_{(y,x)\in G(u|x;\theta)\times\{x\}} g(y,x) dF_\theta(u)\\ &= \int \bar{g}(u) f_\theta(u) du,\end{aligned} \tag{C.3}$$

where $\bar{g}(u) = \max_{(y,x)\in G(u|x;\theta)\times\{x\}} g(y,x)$. This, boundedness of g, and Assumption 4 imply

$$\begin{aligned}\Big|\int g d\nu^*_\theta - \int g d\nu^*_{\theta'}\Big| &= \Big|\int \bar{g}(u)(f_\theta(u) - f_{\theta'}(u))du\Big|\\ &\lesssim \|f_\theta - f_{\theta'}\|_{L^1_\zeta} \lesssim \|\theta - \theta'\|.\end{aligned} \tag{C.4}$$

This ensures (C.1). Showing (C.2) is analogous and is omitted. $\square$

The following proposition shows that the sample log-likelihood converges to its population counterpart uniformly over a set of distributions that are consistent with the null or local alternative hypotheses.

Proposition C.1 (ULLN)**.** *Suppose Assumptions 1-3 hold. Let $h \in \mathcal{V}_1$. Then, for any $\theta_0 \in \Theta_0$ and $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ that does not depend on δ such that*

$$\sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n\Big(\sup_{\delta\in\Theta_\delta} \big|n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - E_{Q_0}[\ln q_{\beta_0,\delta}]\big| \geq \epsilon\Big) < \epsilon,\ \forall n \geq N_\epsilon. \tag{C.5}$$

Proof. Below, let ν_θ and ν^*_θ be a belief function and its conjugate induced by the correspondence $(u,x) \mapsto G(u|x;\theta) \times \{x\}$ on $\mathcal{Y}\times\mathcal{X}$. That is, they are set functions such that

$$\nu_\theta(A) = \int_{\mathcal{X}}\int_{\mathcal{U}} 1\{G(u|x;\theta)) \times \{x\} \subseteq A\} dF_\theta(u) dq_x(x),\ A \subset \mathcal{Y}\times\mathcal{X} \tag{C.6}$$

$$\nu^*_\theta(A) = \int_{\mathcal{X}}\int_{\mathcal{U}} 1\{G(u|x;\theta)) \times \{x\} \cap A \neq \varnothing\} dF_\theta(u) dq_x(x).\ A \subset \mathcal{Y}\times\mathcal{X}. \tag{C.7}$$

A key observation is that, for any $\theta_0 \in \Theta_0$,

$$\int \ln q_{\beta_0,\delta} d\nu_{\theta_0} = E_{Q_{\theta_0}}[\ln q_{\beta_0,\delta}] = \int \ln q_{\beta_0,\delta}(y|x) d\nu^*_{\theta_0}. \tag{C.8}$$

This is because the model is complete under H_0 by Assumption 1 and the fact that the Choquet integrals with respect to ν_{θ_0} and $\nu^*_{\theta_0}$ coincide with each other in such a setting.

Note that one may write the event (i.e. the argument of P^n) in (C.5) as the union of the following two events:

$$A^U_n = \Big\{s^n : \sup_{\delta\in\Theta_\delta}\Big(n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0}\Big) \geq \epsilon\Big\} \tag{C.9}$$

$$A^L_n = \Big\{s^n : \inf_{\delta\in\Theta_\delta}\Big(n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0}\Big) \leq -\epsilon\Big\}. \tag{C.10}$$

Let $K^n = \prod_{i=1}^n K_i$ be a random set whose distribution follows the law induced by $m_{\theta_0+h/\sqrt{n}}$. Below, we simply write $K^n \sim m_{\theta_0+h/\sqrt{n}}$. Note that

$$\begin{aligned}
\sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^U_n \cap A^L_n) &= F_{\theta_0+h/\sqrt{n}}\Big(K^n \cap (A^U_n \cup A^L_n) \neq \varnothing\Big) &\text{(C.11)}\\
&\leq F_{\theta_0+h/\sqrt{n}}(K^n \cap A^U_n \neq \varnothing) + F_{\theta_0+h/\sqrt{n}}(K^n \cap A^L_n \neq \varnothing) &\text{(C.12)}\\
&= \nu^*_{\theta_0+h/\sqrt{n}}\Big(\sup_{\delta\in\Theta_\delta}\big[n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0}\big] \geq \epsilon\Big) &\text{(C.13)}\\
&\quad + \nu^*_{\theta_0+h/\sqrt{n}}\Big(\inf_{\delta\in\Theta_\delta}\big[n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0}\big] \leq -\epsilon\Big). &\text{(C.14)}
\end{aligned}$$

By Assumption 3 and Lemma C.1,

$$\Big|\int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}} - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0}\Big| \leq \frac{C'|h|}{\sqrt{n}}, \tag{C.15}$$

implying there exists $\bar{\eta} > 0$ and $N_{\bar{\eta}}$ such that $\sqrt{n}\sup_{\delta\in\Theta_\delta}\Big(\int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}} - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0}\Big) < \bar{\eta}$ for all $n \geq N_{\bar{\eta}}$. Hence, for all $n \geq N_{\bar{\eta}}$, (C.13) is bounded by

$$\nu^*_{\theta_0+h/\sqrt{n}}\Big(\sup_{\delta\in\Theta_\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big[\ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}}\big] \geq \sqrt{n}\epsilon - \bar{\eta}\Big). \tag{C.16}$$

As we show below, we may apply Lemma C.2 to this quantity. Similarly, by Assumption 3 and Lemma C.1, (C.14) is bounded by

$$\nu^*_{\theta_0+h/\sqrt{n}}\Big(\inf_{\delta\in\Theta_\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big[\ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0+h/\sqrt{n}}\big] \leq -\sqrt{n}\epsilon + \bar{\eta}\Big). \tag{C.17}$$

Now, let $\mathcal{G} \equiv \{g = \ln q_{\beta_0,\delta}, \delta \in \Theta_\delta\}$. Then, by Lemma C.3, the induced family of functions $\mathcal{F}_\mathcal{G}$ defined in (C.20) consists of uniformly bounded and Lipschitz functions. By Theorem 2.7.11 in van der Vaart and Wellner (1996), it follows that

$$N_{[]}(\epsilon\|F_\mathcal{G}\|_{L^2(M)}, \mathcal{F}_\mathcal{G}, L^2(M)) \leq N(\epsilon/2, \Theta_\delta, \|\cdot\|) \leq (2\mathrm{diam}(\Theta_\delta)/\epsilon)^{d_\delta}. \tag{C.18}$$

Therefore, $\mathcal{F}_\mathcal{G}$ satisfies the condition of Lemma C.2. Applying the lemma ensures that (C.16) is bounded by

$$\Big(C_{\mathrm{diam}(\Theta_\delta)}\frac{\sqrt{n}\epsilon - \bar{\eta}}{\sqrt{d_\delta}}\Big)^{d_\delta} e^{-2(\sqrt{n}\epsilon-\bar{\eta})^2}, \tag{C.19}$$

which tends to 0 as $n \to \infty$. (C.17) can be handled similarly. This completes the proof. □

Let S be a Euclidean space. Given a family $\mathcal{G}$ of measurable functions on S and a random set $K : \Omega \mapsto \mathcal{K}(S)$, define a family of measurable functions on $\mathcal{K}(S)$ by

$$\mathcal{F}_\mathcal{G} \equiv \Big\{f : f(K) = \max_{s\in K} g(s),\ g \in \mathcal{G}\Big\}. \tag{C.20}$$

We denote the envelope function of $\mathcal{F}_\mathcal{G}$ by $F_\mathcal{G}$. A class $\mathcal{F}$ of uniformly bounded functions is covered by at most $(\frac{D}{\epsilon})^v$ brackets if for positive constants v and D,

$$N_{[]}(\epsilon\|F\|_{L^2(M)}, \mathcal{F}, L^2(M)) \leq \Big(\frac{D}{\epsilon}\Big)^v,\quad 0 < \epsilon < D, \tag{C.21}$$

The following lemma gives concentration inequalities for the suprema (and infima) of empirical processes under plausibility functions.

Lemma C.2. *Let ν^n be a belief function such that $\nu^n(B) = M^n(K^n \subset A)$ for any $A \in \mathcal{K}(S^n)$. Let $\mathcal{G}$ be a family of uniformly bounded measurable functions on S such that $\mathcal{F}_\mathcal{G}$ in (C.20) is covered by at most $(\frac{D}{\epsilon})^v$ brackets. Then, for all $t > 0$*

$$\nu^{*,n}\Big(\sup_{g\in\mathcal{G}}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big[g(s_i) - \int g d\nu^*\big] \geq t\Big) \leq \Big(C_D\frac{t}{\sqrt{v}}\Big)^v e^{-2t^2}, \tag{C.22}$$

$$\nu^{*,n}\Big(\inf_{g\in\mathcal{G}}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big[g(s_i) - \int g d\nu\big] \leq -t\Big) \leq \Big(C_D\frac{t}{\sqrt{v}}\Big)^v e^{-2t^2}. \tag{C.23}$$

for some C_D that depends on D only.

Proof. Define the following events

$$B_n^U = \Big\{ s^n : \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^* \big] \geq t \Big\} \tag{C.24}$$

$$B_n^L = \Big\{ s^n : \inf_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu \big] \leq -t \Big\}. \tag{C.25}$$

Observe that

$$\begin{aligned} K^n \cap B_n^U \neq \varnothing \;\Leftrightarrow\; & \sup_{s^n \in K^n} \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^* \big] \geq t & \text{(C.26)} \\ \Leftrightarrow\; & \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[\sup_{s_i \in K_i} g(s_i) - \int g d\nu^* \big] \geq t & \text{(C.27)} \\ \Leftrightarrow\; & \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[\sup_{s_i \in K_i} g(s_i) - \int \sup_{s \in K} g(s) dM(K) \big] \geq t & \text{(C.28)} \\ \Leftrightarrow\; & \sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(K_i) - E_M[f(K)] \big] \geq t. & \text{(C.29)} \end{aligned}$$

Therefore,

$$\begin{aligned} \nu^{*,n}\Big(\sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^* \big] \geq t \Big) &= M^n (K^n \cap B_n^U \neq \varnothing) & \text{(C.30)} \\ &= M^n \Big(\sup_{g \in \mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(K_i) - E_M[f(K)] \big] \geq t \Big), & \text{(C.31)} \end{aligned}$$

By Theorem 1.3 (ii) in Talagrand (1994), for all $t > 0$,

$$M^n \Big(\sup_{g \in \mathcal{G}} \big(\frac{1}{\sqrt{n}} \sum_{i=1}^n f(K_i) - E_M[f(K)] \big) \geq t \Big) \leq M^n \Big(\|\mathbb{G}_n f\|_{\mathcal{F}} \geq t \Big) \leq \Big(C_D \frac{t}{\sqrt{v}} \Big)^v e^{-2t^2}. \tag{C.32}$$

A similar argument can be applied to B_n^L as well. □

Let K be a subset of $\mathcal{S} = \mathcal{Y} \times \mathcal{X}$. The following lemma shows that $f_\delta(K) \equiv \max_{(y,x) \in K} \ln q_{\beta_0,\delta}(y|x)$ is uniformly bounded and Lipschitz, which provides a control of the covering number.

Lemma C.3. *Suppose Assumption 3 holds. Then, (i) f_δ is uniformly bounded; and (ii) for any $\delta, \delta' \in \Theta_\delta$,*

$$|f_\delta(K) - f_{\delta'}(K)| \lesssim \|\delta - \delta'\|. \tag{C.33}$$

Proof. (i) follows from the map $\delta \mapsto M(\delta)$ being continuous by Assumption 3 and hence achieving a finite maximum on the compact set Θ_δ.

(ii) Let $s = (y, x)$ and let $g(\delta, s) = \ln q_{\beta_0,\delta}(y|x)$. By Assumption 3,

$$f_{\delta'}(K) = \max_{(y,x) \in K} \Big(g(\delta', s) - g(\delta, s) + g(\delta, s) \Big) \leq \max_{s \in K} \Big(\|\delta - \delta'\| + g(\delta, s) \Big) = f_\delta(K) + \|\delta - \delta'\|. \tag{C.34}$$

Similarly,

$$f_{\delta'}(K) = \max_{s\in K}\Big(g(\delta',s) - g(\delta,s) + g(\delta,s)\Big) \geq \max_{s\in K}\Big(-\|\delta-\delta'\| + g(\delta,s)\Big) = f_\delta(K) - \|\delta-\delta'\|. \tag{C.35}$$

Combining the two inequalities above yields (C.33). □

Below, we write $X_n = O_{P^n}(a_n)$ uniformly in $P^n \in \mathcal{F}_n$ if for any $\epsilon > 0$, there exist finite $M > 0$ and $N > 0$ such that $\sup_{P^n\in\mathcal{F}_n} P^n(|X_n/a_n| > M) < \epsilon$ for all $n > N$.

Theorem C.1. *Suppose Assumptions 1-3 hold. Then, for any $\eta > 0$,*

$$\lim_{n\to\infty} \inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n\Big(\|\hat{\delta}_n - \delta_0\| < \eta\Big) = 1 \tag{C.36}$$

and $\mathbb{M}_n(\hat{\delta}_n) \geq \mathbb{M}_n(\delta_0) - O_{P^n}(r_n^{-2})$ uniformly in $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$.

Proof. The proof is based on the standard argument for the consistency of extremum estimators (see e.g. Newey and McFadden, 1994). A slight difference is that one needs a uniform law of large numbers under any sequence $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$, which is established by Proposition C.1. For each $\delta \in \Theta_\delta$, recall that $\mathbb{M}(\delta) \equiv E_{Q_0}[\ln q_{\beta_0,\delta}]$ and let $\mathbb{M}_n(\delta) \equiv n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i)$. Given any neighborhood V of δ_0, we want to show that $\hat{\delta}_n \in V$, wp$\to$ 1 uniformly over $\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$. For this, it suffices to show that $\inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(\mathbb{M}(\hat{\delta}_n) < \inf_{\delta\in\Theta\cap V^c}\mathbb{M}(\delta)) \to 1$. Let $\epsilon \equiv \inf_{\delta\in\Theta\cap V^c}\mathbb{M}(\delta) - \mathbb{M}(\delta_0)$. This constant is well-defined since $\inf_{\Theta\cap V^c}\mathbb{M}(\delta) = \mathbb{M}(\delta^*) > \mathbb{M}(\delta_0)$ for some $\delta^* \in \Theta \cap V^c$ by Assumption 3 and the compactness of Θ_δ.

Let $A_{1n} \equiv \{\omega : \mathbb{M}(\hat{\delta}_n) < \mathbb{M}_n(\hat{\delta}_n) + \epsilon/3\}, A_{2n} \equiv \{\omega : \mathbb{M}_n(\hat{\delta}_n) < \mathbb{M}_n(\delta_0) + \epsilon/3\}, A_{3n} \equiv \{\omega : \mathbb{M}_n(\delta_0) < \mathbb{M}(\delta_0) + \epsilon/3\}$. For any $\omega \in A_{1n}\cap A_{2n}\cap A_{3n}$,

$$\begin{aligned}\mathbb{M}(\hat{\delta}_n) &< \mathbb{M}_n(\hat{\delta}_n) + \epsilon/3 \\ &< \mathbb{M}_n(\delta_0) + 2\epsilon/3 \\ &< \mathbb{M}(\delta_0) + \epsilon.\end{aligned}$$

Therefore,

$$\inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(\mathbb{M}(\hat{\delta}_n) < \mathbb{M}(\delta_0) + \epsilon) \geq \inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A_{1n}\cap A_{2n}\cap A_{3n}) \tag{C.37}$$

$$\geq \inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}}\Big(1 - P^n(A^c_{1n}) - P(A^c_{2n}) - P(A^c_{3n})\Big) \tag{C.38}$$

$$\geq 1 - \sum_{j=1}^3 \sup_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{jn}). \tag{C.39}$$

Note that, for any h, $\sup_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{1n}) \to 0$ and $\sup_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{3n}) \to 0$ by Proposition C.1. Also note that $\sup_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{2n}) \to 0$ by the construction of $\hat{\delta}_n$. □

Proof of Proposition 3.1. The result follows immediately from Theorem 3.2.5 in van der Vaart and Wellner (1996) with $\phi_n(\zeta) = \zeta$, $r_n = \sqrt{n}$, and applying their argument uniformly over $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$. □

C.2 Size control

Proof of Theorem 3.1. We show the result for the general setting, in which the orthogonalized score is used to construct $\hat{S}_n$. To simplify the exposition below, we assume I_{θ_0} is known for now. Let

$$g_n^*(\beta_0) = C_{\beta_0,n}^* - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^*, \tag{C.40}$$

where

$$C_{\beta_0,n}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta_0), \quad C_{\delta,n}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_\delta(Y_i|X_i;\beta_0,\delta_0). \tag{C.41}$$

By Assumption 1, $P_0^n = Q_{\theta_0}^n \times q_X^n$ for some unique product measure. By Assumption 2 and arguing as in Theorem 7.2 in van der Vaart (2000), we have $E[s_\beta(Y_i|X_i;\beta_0,\delta_0)] = E[s_\delta(Y_i|X_i;\beta_0,\delta_0)] = 0$, where expectation is with respect to P_0. By the square integrability of s_β and s_δ ensured by Assumption 2, the central limit theorem for i.i.d. sequences ensures

$$C_n^* = \begin{bmatrix} C_{\beta_0,n}^* \\ C_{\delta,n}^* \end{bmatrix} \overset{P_0^n}{\rightsquigarrow} N(0, I_{\theta_0}). \tag{C.42}$$

Observing that $\sqrt{n} g_n^*(\beta_0) = [I_{d_\beta}, -I_{\beta,\delta} I_\delta^{-1}] C_n^*$ and applying the continuous mapping theorem, we obtain

$$\sqrt{n} g_n^*(\beta_0) \overset{P_0^n}{\rightsquigarrow} N(0, V_0), \tag{C.43}$$

where $V_0 = I_\beta - I_{\beta,\delta} I_\delta^{-1} I_{\delta,\beta}$. Define

$$S_n^* = n g_n^*(\beta_0)' V_0^{-1} g_n^*(\beta_0) - \inf_{h \in \mathcal{V}_1} n (g_n^*(\beta_0) - h)' V_0^{-1} (g_n^*(\beta_0) - h). \tag{C.44}$$

By (C.43), it then follows that

$$S_n^* \overset{P_0^n}{\rightsquigarrow} S, \tag{C.45}$$

where S is as in (3.27).

For the desired result, it remains to show S_n is asymptotically equivalent to S_n^* under P_0^n. For each δ, let $\mathbb{G}_n s_\beta(\delta) \equiv \frac{1}{n} \sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta) - E[s_\beta(Y_i|X_i;\beta_0,\delta)]$. We may then write

$$\begin{aligned}
\sqrt{n} g_n(\beta_0) - \sqrt{n} g_n^*(\beta_0) &= \mathbb{G}_n s_\beta(\hat{\delta}_n) - \mathbb{G}_n s_\beta(\delta_0) \\
&\quad - \sqrt{n}(E[s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n)] - E[s_\beta(Y_i|X_i;\beta_0,\delta_0)]) \\
&\quad - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n} + I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* \\
&= \mathbb{G}_n s_\beta(\hat{\delta}_n) - \mathbb{G}_n s_\beta(\delta_0) - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* + o_{P^n}(1) \\
&\quad - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n} + I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* \\
&= o_{P^n}(1),
\end{aligned} \tag{C.46}$$

where the last equality follows from the stochastic equicontinuity of $\mathbb{G}_n s_\beta$, $\sqrt{n}$-consistency of $\hat{\delta}_n$, and $C_{\delta,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_\delta(Y_i|X_i;\beta_0,\hat{\delta}_n) = o_{P^n}(1)$ by the first-order condition for the RMLE.

Let $\varphi(x) = x' V_0^{-1} x - \inf_{h \in \mathcal{V}_1} (x-h)' V_0^{-1} (x-h)$. Note that $x \mapsto \inf_{h \in \mathcal{V}_1} (x-h)' V_0^{-1} (x-h)$ is continuous due to Berge's maximum theorem (Aliprantis and Border, 2006, Theorem 17.31). Hence, φ is continuous. By (C.46) and the continuous mapping theorem,

$$\hat{S}_n - S_n^* = \varphi(\sqrt{n} g_n(\beta_0)) - \varphi(\sqrt{n} g_n^*(\beta_0)) = o_{P^n}(1). \tag{C.47}$$

By (C.45) and (C.47),

$$\lim_{n\to\infty} P_0^n(\hat{S}_n > c_\alpha) = \alpha. \tag{C.48}$$

This establishes the claim of the theorem.

Note that we assumed I_{θ_0} was known. In general, it can be consistently estimated by $\hat{I}_n = n^{-1}\sum_{i=1}^n s_\theta(Y_i|X_i;\beta_0,\hat{\delta}_n)s_\theta(Y_i|X_i;\beta_0,\hat{\delta}_n)'$. To see this, let $s_{\theta,j}$ be the j-th component of s_θ. For each j and k, define

$$\xi_{j,k}(Y_i, X_i;\delta) \equiv s_{\theta,j}(Y_i|X_i;\beta_0,\delta)s_{\theta,k}(Y_i|X_i;\beta_0,\delta). \tag{C.49}$$

For the (j,k)-th component of $\hat{I}_n$, we then have

$$\begin{aligned}
[\hat{I}_n]_{j,k} - [I_{\theta_0}]_{j,k} &= \frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i, X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i, X_i;\delta_0)] \\
&= \big(\frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i, X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i, X_i;\hat{\delta}_n)\big) \\
&\quad + \big(E_{q_{\theta_0}}[\xi_{j,k}(Y_i, X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i, X_i;\delta_0)\big) = o_{P^n}(1),
\end{aligned}$$

where the last equality follows because of $\sup_\delta |\frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i, X_i;\delta) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i, X_i;\delta)| = o_{P^n}(1)$ due to Assumption 6, $\hat{\delta}_n$'s consistency by Theorem C.1, and the continuity of $\delta \mapsto E[\xi_{j,k}(Y_i, X_i;\delta)]$. Given this, showing the claim of the theorem with the estimated I_{θ_0} is straightforward by applying Slutsky's theorem.

□