Journal of Econometrics 187 (2015) 1–17

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Model selection tests for moment inequality models

Xiaoxia Shi

University of Wisconsin at Madison, 1180 Observatory Drive, Madison, WI, 53711, United States

ARTICLE INFO

Article history:
Received 4 November 2011
Received in revised form
11 February 2014
Accepted 20 January 2015
Available online 11 February 2015

JEL classification:
C12
C52

Keywords:
Asymptotic size
Kullback–Leibler divergence
Model selection test
Moment inequalities
Overlapping models
Partial identification

ABSTRACT

We propose Vuong-type tests to select between two moment inequality models based on their Kullback–Leibler distances to the true data distribution. The candidate models can be either non-overlapping or overlapping. For each case, we develop a testing procedure that has correct asymptotic size in a uniform sense despite the potential lack of point identification. We show both procedures are consistent against fixed alternatives and local alternatives converging to the null at rates arbitrarily close to $n^{-1/2}$. We demonstrate the finite-sample performance of the tests with Monte Carlo simulation of a missing data example. The tests are relatively easy to implement.

1. Introduction

Models defined by moment inequalities (and possibly some equalities) have gained substantial popularity over recent years as researchers try to move away from ad hoc structural assumptions in various areas of economics.[1] Model selection problems in this context arise naturally when researchers consider more than one economic theory, each generating a set of moment inequalities, or when they consider different parametrizations to form the moment functions. While there is an emerging literature on parameter inference for moment inequality models, a procedure for model selection has not been available.[2] Existing model selection methods for standard models (e.g. Vuong, 1989, Kitamura, 2000, AIC, or BIC) are not applicable because moment inequality models are non-traditional in the ways discussed shortly below.

This paper provides a way to select the better model from two competing moment inequality models. We design quasi-likelihood-ratio tests for the null hypothesis that both models are equally close to the true data distribution in terms of the Kullback–Leibler (KL) divergence. When the null does not hold, the tests direct the researcher to the model that is closer to the true distribution with probability approaching one. Our tests are relatively easy to compute for two reasons. First, they use standard normal critical values. Second, although the sample criterion functions can have multiple (or even a continuum of) maximizers due to partial identification, one does not need to compute all the maximizers to implement the tests.

Moment inequality models are non-traditional in two ways. First, parameters in these models typically are not point-identified. For that reason, the maximizers of a sample criterion function do not converge to a point in the parameter space. Thus, traditional model selection methods that rely on the asymptotic normality of the maximizers do not apply. Second, moment inequality models have slackness parameters whose (pseudo-) true values may be on the boundary of the parameter space.[3] The parameter-on-the-boundary problem makes the criterion function for the original

E-mail address: xshi@ssc.wisc.edu.

[1] They have been used to model discrete games with multiple equilibria (Andrews et al., 2004, Ciliberto and Tamer, 2009), to deal with missing or interval data (Manski, 2005), to study dynamic games that are otherwise too complicated to analyze empirically (Pakes et al., 2007, Pakes, 2010) and to increase the precision of estimators in dynamic macroeconomics models (Moon and Schorfheide, 2009).

[2] A non-exhaustive list of papers on parameter inference of moment inequality models includes Chernozhukov et al. (2007), Andrews and Barwick (2012), Bugni (2010), Canay (2010), Romano and Shaikh (2010), Andrews and Guggenberger (2009), Andrews and Soares (2010) and Andrews and Shi (2013a,b).

[3] One can view the moment inequality model $Em(X_i, \theta) \geq 0$ as a moment equality model with an additional parameter a: $Em(X_i, \theta) - a = 0$. The additional parameter is the slackness parameter. The space of a is $R_+^{d_m}$. The true value of a is on the boundary of $R_+^{d_m}$ whenever a moment inequality holds as an equality under the

http://dx.doi.org/10.1016/j.jeconom.2015.01.004

model parameters non-differentiable even in the limit. The non-differentiability can occur anywhere in the original parameter space. Thus, the first-order-condition method or the standard quadratic approximation method cannot be used to derive the convergence rate of the estimators.

The first nontraditional feature prompts us to develop a new technique utilizing the stochastic equicontinuity of certain empirical processes to show the asymptotic normality of the quasi-likelihood ratio statistic and the consistency of an estimator of its asymptotic variance. The technique does not require any convergence rate of the sample maximizers. We only need a weak notion of consistency: the sample maximizers approach the pseudo-true set as the sample size goes to infinity. This technique potentially is useful to establish the asymptotic distribution of the Vuong (1989) test statistic in parametric models and moment equality models as well when the Hessian matrix of the likelihood ratio is not invertible.

The asymptotic normality and the consistency results mentioned above are sufficient for developing a valid model selection test if the asymptotic variance of the quasi-likelihood ratio statistic is bounded away from zero. The latter condition holds when the two models compared are non-overlapping in the sense defined in latter sections. When the two models are overlapping, the convergence rate of the sample maximizers is needed.

The second nontraditional feature of moment inequality models made the traditional approaches to derive convergence rate not applicable. We modify the standard quadratic approximation method and construct quadratic upper and lower bounds for the sample and population criterion functions. Combining those bounds, we show that the sample maximizers approach the pseudo-true set at $n^{-1/2}$-rate. The rate is then used to motivate an adjustment factor to the studentized quasi-likelihood ratio statistic. The adjustment factor guarantees that the adjusted test is uniformly valid for overlapping models.

The tests proposed in this paper extend the Vuong test (for maximum likelihood models) proposed in the seminal paper Vuong (1989) to models defined by moment inequalities. As such, this paper belongs to the literature that extends Vuong (1989) to various other types of models. Kitamura (2000) and Rivers and Vuong (2002) extend the Vuong test to models defined by moment *equalities*. In particular, Kitamura (2000) employs exponential tilting criterion, which is adapted to moment inequality models in the current paper. Chen et al. (2007) propose a Vuong-type procedure to select between a parametric model and a moment equality model. All these previous papers assume that the true parameters are point-identified and are in the interior of the parameter space. These assumptions are suitable for parametric models and moment equality models, but not for the moment inequality models considered in this paper. On the other hand, this paper does not make those assumption. Thus, our tests apply to point or partially identified moment inequality or equality models with or without parameter on the boundary. In the special case of non-overlapping point identified moment equality models without parameter on the boundary, our test is the same as Kitamura's (2000).

In addition to addressing the partial identification and parameter-on-the-boundary problems, another important feature distinguishing our tests from the other Vuong-type tests is that we choose the critical values based on uniform asymptotics which guarantee correct asymptotic sizes of the tests. Vuong-type tests with critical values chosen based on pointwise asymptotics may have size distortion when the candidate models are overlapping. The reason is that the pointwise asymptotic distributions of the test statistics are discontinuous in the data generating process. When the data generating process is close to the discontinuity point, the finite sample distributions of the test statistics are not well approximated by their pointwise asymptotic distributions. The poor approximation causes size distortion in finite samples (Shi, forthcoming). We adjust the test statistic in the overlapping case to take into account the discontinuity and by doing so control the asymptotic size of the tests uniformly.

An alternative to our Vuong-type framework is the Cox (1961)-type nonnested hypothesis testing framework. For a Cox-type test, the null hypothesis is that a model $\mathcal{P}$ is correctly specified and the alternative hypothesis is that an alternative model $\mathcal{Q}$ is correctly specified. Though frequently used to choose one model from multiple candidate models, Cox-type tests are intended as a procedure for model evaluation rather than model selection. A Cox-type test does not have a clear interpretation when both models are misspecified. For details on Cox-type tests, see the seminal paper by Cox (1961), the survey papers by Gourieroux and Monfort (1994) and Pesaran and Weeks (1999), generalizations to the encompassing principle by Mizon and Richard (1986), and the extension to moment equality models by Ramalho and Smith (2002). It is of interest to extend the moment encompassing principle to partially-identified moment inequality models possibly using some of the techniques developed in this paper. We leave this to a separate project.

The rest of the paper is organized as follows. Section 2 introduces the model selection problem for moment inequality models and gives a few examples. Section 3 presents preliminaries on the pseudo-distance measure and the solution to the distance-minimizing problem. Section 4 describes the tests, one for non-overlapping models and the other for overlapping models. Sections 5 and 6 establish the asymptotic size of the test for non-overlapping models and that for overlapping models, respectively. Section 7 determines the power properties of the tests. Section 8 presents Monte Carlo simulation results for a missing data example. The proofs are in the appendix.

We use $N_\delta(\theta)$ to denote a closed ball centered at θ with radius δ, $\|\cdot\|$ to denote the Euclidean norm, and "$\ll$" to denote "is absolutely continuous with respect to (w.r.t., hereafter)". We use X_i to denote an observation, $\mathcal{X}$ to denote the space on which X_i is defined. We use $\mathcal{P}$ and $\mathcal{Q}$ to denote the candidate models, and P and Q to denote generic distributions in $\mathcal{P}$ and $\mathcal{Q}$, respectively. We use μ to denote a generic true distribution on $\mathcal{X}$, which does not necessarily belong to either of the models. We use Greek letters θ and β to denote the finite-dimensional parameters in the models, Θ and B to denote the corresponding parameter spaces, and m and g to denote the moment functions.

2. Model selection problems

We consider two moment inequality/equality models $\mathcal{P} = \bigcup_{\theta\in\Theta}\mathcal{P}_\theta$ and $\mathcal{Q} = \bigcup_{\beta\in B}\mathcal{Q}_\beta$, where $\mathcal{P}_\theta$ and $\mathcal{Q}_\beta$ are the set of distributions that are consistent with the moment conditions for parameters θ and β, respectively:

$$\mathcal{P}_\theta = \left\{ P: \begin{array}{l} E_P m_j(X_i,\theta) = 0 \text{ for } j = 1,\ldots,d_p, \\ E_P m_j(X_i,\theta) \geq 0 \text{ for } j = d_p+1,\ldots,d_m \end{array} \right\}$$

$$\mathcal{Q}_\beta = \left\{ Q: \begin{array}{l} E_Q g_j(X_i,\beta) = 0 \text{ for } j = 1,\ldots,d_q, \\ E_Q g_j(X_i,\beta) \geq 0 \text{ for } j = d_q+1,\ldots,d_g \end{array} \right\}. \quad (2.1)$$

In the above equation, $\{X_i \in \mathcal{X}\}_{i=1}^n$ is a random sample generated from μ, $m = (m_1,\ldots,m_{d_p},m_{d_p+1},\ldots,m_{d_m})'$ and $g = (g_1,\ldots,g_{d_q},g_{d_q+1},\ldots,g_{d_g})'$ are R^{d_m} and R^{d_g}-valued moment functions known up to the finite-dimensional parameters θ and β, respectively, $\Theta \subset R^{d_\theta}$, $B \subset R^{d_\beta}$, and E_P denotes the expectation

true data distribution. In this example, $\{X_i\}$ is the data, m is a R^{d_m}-valued moment function and θ is a finite-dimensional parameter.

under the distribution P. Either model can be over, just or under-identified, that is, d_p or d_m (d_q or d_g) can be smaller than, larger than, or equal to d_θ (d_β). The true distribution μ may or may not belong to either model. Model $\mathcal{P}$ is called **correctly specified** if $\mu \in \mathcal{P}$ and is called **misspecified** otherwise.

The goal of this paper is to compare models $\mathcal{P}$ and $\mathcal{Q}$ and select the one that is closer to the true distribution μ in terms of a pseudo-distance measure. Let $d(P, \mu)$ be a pseudo-distance between a distribution P and μ. The pseudo distance from a model $\mathcal{P}$ to μ is defined by $d(\mathcal{P}, \mu) = \inf_{P \in \mathcal{P}} d(P, \mu)$. We want to construct model selection tests for the null hypothesis

$$H_0 : d(\mathcal{P}, \mu) = d(\mathcal{Q}, \mu). \tag{2.2}$$

The choice of d is discussed in the next section.

Now, we give a few illustrative examples of model selection problems in the context of moment inequalities. Special cases of Example 1 are studied in the Monte Carlo section (Section 8).

Example 1 (*Interval Outcome in Regression Models*)**.** Consider the regression models with interval outcomes in Manski (2005). A model selection problem of potential interest is selecting different regressors or functional forms for the regression functions. Let Y be a latent random variable (e.g. wealth) that is not perfectly observed. Only an upper bound, $\overline{Y}$, and a lower bound, $\underline{Y}$, on Y are observed. Let X be a vector of explanatory variables and $Y = r(X, \theta) + \varepsilon$, where r is a function known up to a finite-dimensional parameter θ. Let Z be a vector of potential instrument variables such that $E(\varepsilon \cdot I(Z)) = 0$ for some positive (vector-valued) function I of Z. Then, the models $\mathcal{P} = \cup_{\theta\in\Theta} \mathcal{P}_\theta$ and $\mathcal{Q} = \cup_{\beta\in B} \mathcal{Q}_\beta$ where

$$\mathcal{P}_\theta = \{P : E_P[(\overline{Y} - r_1(X, \theta))I(Z)] \ge 0 \;\&\; E_P[(r_1(X, \theta) - \underline{Y})I(Z)] \ge 0\}$$

$$\mathcal{Q}_\beta = \{Q : E_Q[(\overline{Y} - r_2(X, \beta))I(Z)] \ge 0 \;\&\; E_Q[(r_2(X, \beta) - \underline{Y})I(Z)] \ge 0\}, \tag{2.3}$$

where r_1 and r_2 are two regression functions. Note that the distributions P and Q are defined on the space of the **observed** random variables $(\overline{Y}, \underline{Y}, X, Z)$.

Another model selection problem arises when one considers a different choice of instruments. The formulation of the competing models is similar to (2.3), except that r_1 and r_2 are the same and we have I_1 instead of I in model $\mathcal{P}$ and I_2 in model $\mathcal{Q}$.

Example 2 (*Interval Regressor in Regression Models*)**.** Consider the regression models with interval regressors in Manski and Tamer (2002). Let Y be a continuous dependent variable, v be a regressor that is not observed perfectly but in intervals $[\underline{v}, \overline{v}]$. Let X represent other regressors. Assume that $E(Y|X, v) = f(x, v, \theta)$, where f is a function known up to the finite-dimensional parameter θ. As in Manski and Tamer (2002), if we assume that f is weakly *increasing* in v, we obtain the moment inequality model $\mathcal{P} = \cup_{\theta\in\Theta} \mathcal{P}_\theta$, where

$$\begin{aligned}\mathcal{P}_\theta = \{P : &E_P[(Y - f(X, \underline{v}, \theta))I(X, \underline{v}, \overline{v})] \ge 0\\ &\& \; E_P[(f(X, \overline{v}, \theta) - Y)I(X, \underline{v}, \overline{v})] \ge 0\},\end{aligned} \tag{2.4}$$

where $I(X, \underline{v}, \overline{v})$ can be any vector of positive instrument functions.[4] On the other hand, if we assume that f is weakly *decreasing* in v, we have a different moment inequality model $\mathcal{Q} = \cup_{\beta\in B} \mathcal{Q}_\beta$, where

$$\begin{aligned}\mathcal{Q}_\beta = \{Q : &E_Q[(f(X, \underline{v}, \beta) - Y)I(X, \underline{v}, \overline{v})] \ge 0\\ &\& \; E_Q[(Y - f(X, \overline{v}, \beta))I(X, \underline{v}, \overline{v})] \ge 0, \beta \in B\}.\end{aligned} \tag{2.5}$$

By comparing models $\mathcal{P}$ and $\mathcal{Q}$, one can determine which sign assumption on $\partial f/\partial v$ is more consistent with the data.

Example 3 (*Entry Game — Cross-firm Effect*)**.** Consider the entry game example discussed in Tamer (2003), Andrews et al. (2004) and Ciliberto and Tamer (2009). Consider a 2×2 version with the following payoff matrix:

		Firm 2	
		0	1
Firm 1	0	0, 0	$0, X_2'\theta_2 - \varepsilon_2$
	1	$X_1'\theta_1 - \varepsilon_1, 0$	$X_1'\theta_1 + a_1 - \varepsilon_1, X_2'\theta_2 + a_2 - \varepsilon_2$

The observable random variables are the market characteristics $X \equiv (X_1, X_2)'$ and the game outcome Y. The variable Y may take four values: (0, 0), (0, 1), (1, 0) and (1, 1), where the two numbers in the parenthesis are the equilibrium actions of firm 1 and firm 2, respectively. The coefficients θ_1 and θ_2 are the marginal effects of the characteristics X on profits, and ε_1 and ε_2 are profit shocks unobservable to the econometrician. The parameters a_1 and a_2 are the cross-firm effects, which are the effects of the firms on their opponents' profit when they form a duopoly.

Let $F_{\varepsilon_1,\varepsilon_2}(\cdot, \cdot; \theta_\varepsilon)$ denote the joint c.d.f. of ε_1 and ε_2, $F_{\varepsilon_1}(\cdot; \theta_\varepsilon)$ the marginal c.d.f. of ε_1, and $F_{\varepsilon_2}(\cdot; \theta_\varepsilon)$ the marginal c.d.f. of ε_2. The c.d.f.s are known to the econometrician up to the finite-dimensional parameter θ_ε. Assume that the firms have full information about their own and their opponents' payoffs and play a simultaneous-move Nash game. Andrews et al. (2004) assume $a_1 \le 0$ and $a_2 \le 0$ and obtain the moment inequality model $\mathcal{P} = \cup_{\theta\in\Theta} \mathcal{P}_\theta$, where

$$\begin{aligned}\mathcal{P}_\theta = \{P : \; &E_P[(p_j(X, \theta) - 1(Y = j))I(X)] = 0, \text{for } j = (0, 0) \text{ or } (1, 1)\\ &E_P[(p_j(X, \theta) - 1(Y = j))I(X)] \ge 0, j = (0, 1), \text{ or } (1, 0)\},\end{aligned} \tag{2.6}$$

$\theta \equiv (\theta_1', \theta_2', a_1, a_2, \theta_\varepsilon')'$, $I(X)$ is a vector of positive instrument functions, and

$$\begin{aligned}p_{(0,0)}(X, \theta) \; &= \; 1 - F_{\varepsilon_1}(X_1'\theta_1; \theta_\varepsilon) - F_{\varepsilon_2}(X_2'\theta_2; \theta_\varepsilon)\\ &\quad + F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1, X_2'\theta_2; \theta_\varepsilon)\\ p_{(0,1)}(X, \theta) &= F_{\varepsilon_2}(X_2'\theta_2; \theta_\varepsilon) - F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1 + a_1, X_2'\theta_2; \theta_\varepsilon)\\ p_{(1,0)}(X, \theta) &= F_{\varepsilon_1}(X_1'\theta_1; \theta_\varepsilon) - F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1, X_2'\theta_2 + a_2; \theta_\varepsilon)\\ p_{(1,1)}(X, \theta) &= F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1 + a_1, X_2'\theta_2 + a_2; \theta_\varepsilon).\end{aligned} \tag{2.7}$$

On the other hand, if we assume $a_1 \ge 0$ and $a_2 \ge 0$, we obtain a different model $\mathcal{Q} = \cup_{\beta\in B} \mathcal{Q}_\beta$, where

$$\begin{aligned}\mathcal{Q}_\beta = \{Q : \; &E_Q[(p_j(X, \beta) - 1(Y = j))I(X)] \ge 0, \text{for } j = (0, 0) \text{ or } (1, 1)\\ &E_Q[(p_j(X, \beta) - 1(Y = j))I(X)] = 0, j = (0, 1), \text{ or } (1, 0)\},\end{aligned} \tag{2.8}$$

$\beta \equiv (\theta_1', \theta_2', a_1, a_2, \theta_\varepsilon')'$ and p_j for $j = (0, 0), (1, 1), (0, 1)$ and $(1, 0)$ are defined in (2.7).

In some industries, for example the shopping center industry studied in Vitorino (2012), the sign of the cross-firm effect is uncertain. A model selection test comparing the two models above can determine which sign of the cross-firm effects is more consistent with the data.

Example 4 (*Entry Game — Testing Equilibrium Selection Mechanism*)**.** Instead of being agnostic about the equilibrium selection mechanism, one can also specify such a mechanism, as done in Tamer (2003) among others. For example, in the case of negative cross-firm effects, one can assume that the probability of (1, 0) is $H(X, \gamma)$ in case of multiple equilibria. That yields a moment equality model:

$$\begin{aligned}\mathcal{P}_2 = \{P : \; &E_P[(p_j(X, \theta) - 1(Y = j))I(X)] = 0, \text{for } j = (0, 0) \text{ or } (1, 1)\\ &E_P[(p_j(X, \theta) - p_m(X, \theta)H(X, \gamma) - 1(Y = j))I(X)] = 0, j = (0, 1)\\ &\text{for some } (\theta, \gamma) \in \Theta \times \Gamma\},\end{aligned} \tag{2.9}$$

where $p_m(X, \theta) = F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1, X_2'\theta_2; \theta_\varepsilon) - F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1 - a_1, X_2'\theta_2; \theta_\varepsilon) - F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1, X_2'\theta_2 - a_2; \theta_\varepsilon) + F_{\varepsilon_1,\varepsilon_2}(X_1'\theta_1 - a_1, X_2'\theta_2 - a_2; \theta_\varepsilon)$ is the probability that multiple equilibria occur.

[4] Note that the probability measure P's are defined on the space of $(Y, X, \overline{v}, \underline{v})$.

The equilibrium selection rule $H(X, \gamma)$ can be flexibly specified. But even then, it imposes the fundamental assumption that equilibrium selection only depends on observables. A model selection test between $\mathscr{P}$ and $\mathscr{P}_2$ can help to determine whether this assumption is consistent with the data. In this example, the two models are nested.

Example 5 (*Entry Game — Choosing Information Structure*)**.** Model selection test also can be used to choose the information structure of a game-theoretical model. Berry and Tamer (2006) show that the entry game described in Example 3 can be modeled by a different set of moment inequalities, if we assume that the firms do not know their competitors' idiosyncratic profits $(\varepsilon_1, \varepsilon_2)$ but have beliefs about the distributions of $(\varepsilon_1, \varepsilon_2)$. By comparing the new moment inequality model to $\mathscr{P}$ (or $\mathscr{Q}$) in Example 3, one can determine which information structure is more appropriate.

3. Preliminaries on the pseudo-distance measure

There are many possible choices of pseudo-distances on the space of probability distributions. One may prefer one distance to another in a specific problem. Since we deal with a generic problem, we choose the Kullback–Leibler (KL) divergence. The KL divergence from P to μ is

$$d(P, \mu) = \begin{cases} \displaystyle\int p_\mu \log p_\mu d\mu & \text{if } P \ll \mu \\ \infty & \text{otherwise,} \end{cases} \tag{3.1}$$

where p_μ is the density of P with respect to μ.[5] The pseudo-distance above also is called the I-divergence, or the relative entropy of P to μ. For moment condition models, I-divergence motivates the exponential tilting estimation (Kitamura and Stutzer, 1997).

The rest of the discussions in this section – with the exclusion of the formal assumptions and lemmas – are in terms of model $\mathscr{P}$, but they apply to model $\mathscr{Q}$ as well.

In order to measure the distance from the model to the true distribution, one needs to solve the minimization problem $\inf_{P\in\mathscr{P}} d(P, \mu)$. The problem is solved in two steps:

$$\inf_{P\in\mathscr{P}} d(P, \mu) = \inf_{\theta\in\Theta} \inf_{P\in\mathscr{P}_\theta} d(P, \mu), \tag{3.2}$$

where $\mathscr{P}_\theta$ is defined in (2.1). The first step $\inf_{P\in\mathscr{P}_\theta} d(P, \mu)$ is an infinite dimensional minimization problem and can be solved through a finite-dimensional dual problem. The second step is a finite-dimensional minimization problem which may have multiple solutions because model $\mathscr{P}$ may be partially-identified. We discuss both steps in the following subsections.

3.1. The dual problem

The first step minimization $\inf_{P\in\mathscr{P}_\theta} d(P, \mu)$ has a unique solution, if the solution exists. The reason is that $d(P, \mu)$ is strictly convex in P and the set $\mathscr{P}_\theta$ is defined by constraints linear in P and thus is convex. We follow Csiszár (1975) and call the solution to $\inf_{P\in\mathscr{P}_\theta} d(P, \mu)$ the I-**projection** of μ on $\mathscr{P}_\theta$. Denote the I-projection as $P^*_{\mu,\theta}$. For models defined by equality constraints, Csiszár (1975) gives sufficient conditions for the existence of $P^*_{\mu,\theta}$ and shows that $\inf_{P\in\mathscr{P}_\theta} d(P, \mu)$ has a finite-dimensional dual problem under those conditions. We adapt Csiszár's (1975) approach to the context of moment inequality models.

We introduce some notation first. For a data distribution μ, define the dual criterion functions as

$$\begin{aligned} &\mathscr{M}_\mu(\gamma, \theta) := E_\mu \exp\big(\gamma' m(X_i, \theta)\big) \quad \text{and} \\ &\mathscr{N}_\mu(\lambda, \beta) := E_\mu \exp\big(\lambda' g(X_i, \beta)\big). \end{aligned} \tag{3.3}$$

Let the Lagrange multipliers for each θ and β be

$$\begin{aligned} &\gamma^*_\mu(\theta) = \arg \min_{\gamma\in R^{dp}_\infty \times R^{dm-dp}_{+,\infty}} \mathscr{M}_\mu(\gamma, \theta), \quad \text{and} \\ &\lambda^*_\mu(\beta) = \arg \min_{\lambda\in R^{dq}_\infty \times R^{dg-dq}_{+,\infty}} \mathscr{N}_\mu(\lambda, \beta), \end{aligned} \tag{3.4}$$

where $R_\infty = R \cup \{\infty, -\infty\}$ and $R_{+,\infty} = R_+ \cup \{\infty\}$. For every $\theta \in \Theta$, $\gamma^*_\mu(\theta)$ is uniquely defined under Assumption 1(a).

Assumption 1. (a) For all $\theta \in \Theta$, $E_\mu\|m(X_i, \theta)\|^2 < \infty$ and $E_\mu[m(X_i, \theta)m(X_i, \theta)']$ is positive definite,

(b) for all $\theta \in \Theta$, $\|\gamma^*_\mu(\theta)\| < \infty$, and

(c) parts (a)–(b) hold with g, β and λ in place of m, θ and γ.

Although Assumption 1(a) is not a standard assumption in the moment inequality literature, it is standard in the (generalized) empirical likelihood literature and is imposed in other model selection test papers based on generalized empirical likelihood, for example, Kitamura (2000). In the context of moment inequalities, Canay (2010) also imposes this assumption in order to apply the empirical likelihood approach. Assumption 1(b) requires the model not to be too misspecified. A sufficient condition for Assumption 1(b) that is easier to verify is Assumption 1(b)* below.[6]

Assumption 1(b)*. For all $\theta \in \Theta$ and all $\gamma \in R^{dp} \times R^{dm-dp}_+$, $\Pr_\mu(\gamma' m(X_i, \theta) > 0) > 0$.

To show the sufficiency, let $\gamma := (\gamma_1, \ldots, \gamma_{dm})'$ be an arbitrary element in $(R^{dp}_\infty \times R^{dm-dp}_{+,\infty})$ such that $\|\gamma\| = \infty$. Let $\gamma^0 := (\gamma^0_1, \ldots, \gamma^0_{d_m})$ where $\gamma^0_j = 1\,(\gamma_j = \infty) - 1\,(\gamma_j = -\infty)$, and $\gamma^1 := (\gamma^1_1, \ldots, \gamma^1_{d_m})'$ where $\gamma^1_j = \gamma_j \cdot 1(\gamma_j \in R)$. By Assumption 1(b)*, $p^0 := \Pr_\mu(\gamma^{0,\prime} m(X_i, \theta) > 0) > 0$. But $\gamma' m(x, \theta) = \infty \times \gamma^{0,\prime} m(x, \theta) + \gamma^{1,\prime} m(x, \theta)$, which implies that $\gamma' m(x, \theta) = \infty$ if $\gamma^{0,\prime} m(x, \theta) > 0$.[7] Thus, $\Pr_\mu(\gamma' m(X_i, \theta) = \infty) \geq p^0 > 0$. This implies that $\Pr_\mu(\exp(\gamma' m(X_i, \theta)) = \infty) \geq p_0 > 0$. Therefore, $E_\mu \exp(\gamma' m(X_i, \theta)) \geq p_0 \times \infty = \infty$ for the γ's that have infinite norm. Now notice that $E_\mu \exp(\gamma^*_\mu(\theta)' m(X_i, \theta)) := \min_{\gamma\in R^{dp}_\infty \times R^{dm-dp}_{+,\infty}} E_\mu \exp(\gamma' m(X_i, \theta)) \leq E_\mu \exp(0' m(X_i, \theta)) = 1$, where the second inequality holds because $0 \in R^{dp}_\infty \times R^{dm-dp}_{+,\infty}$. This implies that $\gamma^*_\mu(\theta)$ cannot have infinite norm, that is, Assumption 1(b) holds.

Lemma 1 establishes that $\inf_{P\in\mathscr{P}_\theta} d(P, \mu)$ is attained and can be solved through a finite-dimensional dual problem under Assumption 1.

[5] Note that the KL divergence is directional, that is $d(P, \mu) \neq d(\mu, P)$. This makes our hypothesis different from that in Vuong (1989), which is based on $d(\mu, P)$. The duality results in this section are specific to our KL-divergence, but if one assumes the duality as given, the test we develop later in Section 4 can be extended with ease to the KL-divergence of the reversed direction, as well as to generalized empirical likelihood distance measures. For brevity, we do not carry out the generalization, but note that the general distance measure is used in Hsu and Shi (2013) in the context of conditional moment inequalities.

[6] Assumption 1(b)* is violated, for example, when the model is $\mathscr{P} = \{P : E_P(X_{1,i} - \theta) \geq 0, E_P(\theta - X_{2,i}) \geq 0\}$, and $X_{1,i} < X_{2,i}$ a.s. $[\mu]$. To check, let $a = (1, 1)'$. Then, $\Pr_\mu(a' m(X_i, \theta) > 0) = \Pr_\mu(X_{1,i} - X_{2,i} > 0) = 0$.

[7] Here we define $\infty \cdot 0 = 0$.

Lemma 1. *Suppose Assumption 1 holds. Then,*

(a) *for all $\theta \in \Theta$, the I-projection, $P^*_{\mu,\theta}$, of μ on $\mathcal{P}_\theta$ exists and its density w.r.t. μ is*

$$p^*_{\theta,\mu}(x) = \exp\big(\gamma^*_\mu(\theta)' m(x,\theta)\big) / \mathcal{M}_\mu\big(\gamma^*_\mu(\theta), \theta\big),$$

(b) *for all $\theta \in \Theta$, $d(\mathcal{P}_\theta, \mu) = -\log[\mathcal{M}_\mu(\gamma^*_\mu(\theta), \theta)]$,*

(c) *parts* (a)–(b) *hold with g, β, λ, Q, $\mathcal{Q}$ and $\mathcal{N}$ in place of m, θ, γ, P, $\mathcal{P}$ and $\mathcal{M}$.*

3.2. The pseudo-true set and the pseudo-true distribution

The second step infimum in (3.2), $\inf_{\theta\in\Theta} d(\mathcal{P}_\theta, \mu)$, is attained if $d(\mathcal{P}_\theta, \mu)$ is continuous in θ and Θ is compact. These are guaranteed by Assumption 2.

Assumption 2. (a) The parameter spaces Θ and B are compact, and

(b) with probability one, $m(X_i, \cdot)$ and $g(X_i, \cdot)$ are continuous in Θ and B, respectively.

Lemma 2 shows that the infimum $\inf_{\theta\in\Theta} d(\mathcal{P}_\theta, \mu)$ is attained and has a saddle-point dual representation.

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then,*

(a) *there exists a $\theta^* \in \Theta$ such that $\mathcal{M}_\mu(\gamma^*_\mu(\theta^*), \theta^*) = \sup_{\theta\in\Theta} \mathcal{M}_\mu(\gamma^*_\mu(\theta), \theta)$,*

(b) $d(\mathcal{P}, \mu) = -\log\left[\max_{\theta\in\Theta} \min_{\gamma\in R^{d_p}\times R_+^{d_m-d_p}} \mathcal{M}_\mu(\gamma, \theta)\right]$, *and*

(c) *parts* (a)–(b) *hold with g, β, λ, q, Q, $\mathcal{Q}$ and $\mathcal{N}$ in place of m, θ, γ, p, P, $\mathcal{P}$ and $\mathcal{M}$.*

Remark. The function $\gamma^*_\mu(\theta)$ usually has kinks because of the non-negativity constraints in the minimization problem that defines it. This reflects the parameter-on-the-boundary problem discussed in the introduction. At the kinks, $\gamma^*_\mu(\theta)$ is not differentiable in θ. The kinks can occur anywhere in Θ. Thus, the population criterion function, $\mathcal{M}_\mu(\gamma^*_\mu(\cdot), \cdot)$ is non-differentiable.

Because model $\mathcal{P}$ can be partially-identified, $\mathcal{M}_\mu(\gamma^*_\mu(\theta), \theta)$ can have multiple maximizers. We call the set of maximizers the **pseudo-true set:**

$$\Theta^*_\mu = \arg\max_{\theta\in\Theta} \mathcal{M}_\mu(\gamma^*_\mu(\theta), \theta). \tag{3.5}$$

The concept of "pseudo-true set" is generalized from the "pseudo-true parameter" concept in the literature of misspecified point-identified models. The prefix "pseudo" signifies the possibility that the model may be misspecified.

Similarly, we call the distributions that achieve $\min_{P\in\mathcal{P}} d(P, \mu)$ the pseudo-true distributions of model $\mathcal{P}$ under μ. Lemma 1 implies that the set of all pseudo-true distributions of model $\mathcal{P}$ under μ equals $\mathcal{P}^*_\mu := \{P^*_{\theta,\mu} : \theta \in \Theta^*_\mu\}$. This set needs not be a singleton in general (i.e., the pseudo-true distribution might not be unique), but it is guaranteed to be a singleton in these important cases:

(i) $\mu \in \mathcal{P}$. In this case, $\mathcal{P}^*_\mu = \{\mu\}$. This is simply because $d(\mu, \mu) = 0$ and $d(P, \mu) > 0$ for any $P \neq \mu$ by the property of the pseudo-true distance d. Notice that in this case the pseudo-true set Θ^*_μ can still contain multiple values.

(ii) $\mu \notin \mathcal{P}$, but Θ^*_μ is a singleton. This is a natural assumption if the moment equality/inequality model contains no fewer equality restrictions than the number of parameters. For example, our Examples 3 and 4 falls into this scenario if the dimension of $I(X)$ is at least half of the dimension of θ.

(iii) $\mu \notin \mathcal{P}$, but the moment function $m(X_i, \theta)$ depends on θ only through a lower dimensional function of θ: $\beta = b(\theta)$, and $\{b(\theta) : \theta \in \Theta^*_\mu\}$ is a singleton. In other words, if the partial identification is only caused by over-parametrization, the pseudo-true distribution, which does not depend on parametrization, is unique.

The uniqueness of the pseudo-true distribution combined with Lemma 1(a) implies that

$$\gamma^*_\mu(\theta)' m(X_i, \theta) = \gamma^*_\mu(\theta^*)' m(X_i, \theta^*)$$
$$\text{a.s. } [\mu] \text{ for all } \theta, \theta^* \in \Theta^*_\mu. \tag{3.6}$$

Eq. (3.6) is crucial for the quasi-likelihood ratio statistic defined later to be asymptotically normal under H_0. Thus, we maintain the following assumption for data distributions μ that satisfy the null hypothesis (2.2).

Assumption 3. The pseudo-true distributions, P^*_μ and Q^*_μ, of models $\mathcal{P}$ and $\mathcal{Q}$, respectively, are unique under μ.

Remark. The assumption will only be imposed for μ under H_0 and will *not* be imposed under the alternative hypothesis. This makes it relatively weak.[8] In fact, based on the discussion above, this assumption is guaranteed to hold under H_0 in the following important testing scenarios:

(i) $\mathcal{P}$ and $\mathcal{Q}$ are nested and the correct specification of the nesting model is maintained. In standard models, researchers are explicitly or implicitly in this testing scenario whenever the textbook likelihood ratio test with a chi-squared critical value is used. Thus, we believe this is a typical nested testing scenario.

(ii) $\mathcal{P}$ and $\mathcal{Q}$ are nonnested, but the econometrician has the prior knowledge that one of them is correctly specified. Then under H_0, both are correctly specified and hence the pseudo-true distributions are unique.

(iii) Both models are point identified (Θ^*_μ and B^*_μ are singleton sets), which is plausible when both models contain enough number of equality restrictions.

(iv) Partial identification of both models can be reduced to point identification by reparametrization.

4. Model selection tests

In this section we introduce the test statistics first. Then, we formally define non-overlapping models and overlapping models and discuss how the relationship between candidate models affects the asymptotic distributions of the test statistics. Finally, we describe the model selection tests.

4.1. Test statistics

We define the test statistics in this section and give informal discussions on the asymptotics in order to introduce the tests. First, observe that, by Lemma 2(b), the null (2.2) can be written as

$$H_0 : \max_{\theta\in\Theta} \mathcal{M}_\mu(\gamma^*_\mu(\theta), \theta) = \max_{\beta\in B} \mathcal{N}_\mu(\lambda^*_\mu(\beta), \beta). \tag{4.1}$$

The test statistics are based on the sample analogue of the above quantities.

[8] In Supplemental Appendix E, we discuss how to remove this already weak assumption completely using a sample-splitting technique.

Let the sample criterion functions be

$$\begin{aligned}\widehat{\mathcal{M}}_n(\gamma,\theta) &= n^{-1}\sum_{i=1}^{n}\exp(\gamma' m(X_i,\theta)) \quad \text{and}\\ \widehat{\mathcal{N}}_n(\lambda,\beta) &= n^{-1}\sum_{i=1}^{n}\exp(\lambda' g(X_i,\beta)).\end{aligned} \tag{4.2}$$

Let the sample saddle points be

$$\begin{aligned}\hat{\gamma}_n(\theta) &= \arg\min_{\gamma\in R^{dp}\times R_+^{dm-dp}} \widehat{\mathcal{M}}_n(\gamma,\theta),\\ \hat{\lambda}_n(\beta) &= \arg\min_{\lambda\in R^{dq}\times R_+^{dg-dq}} \widehat{\mathcal{N}}_n(\lambda,\beta),\\ \widehat{\Theta}_n &= \arg\max_{\theta\in\Theta} \widehat{\mathcal{M}}_n(\hat{\gamma}_n(\theta),\theta), \quad \text{and}\\ \widehat{B}_n &= \arg\max_{\beta\in B} \widehat{\mathcal{N}}_n(\hat{\lambda}_n(\beta),\beta),\end{aligned} \tag{4.3}$$

where $\widehat{\Theta}_n$ and $\widehat{B}_n$ are not necessarily singletons.

We use the quasi-likelihood ratio (QLR) statistic:

$$\widehat{QLR}_n = \max_{\theta\in\Theta} \widehat{\mathcal{M}}_n(\hat{\gamma}_n(\theta),\theta) - \max_{\beta\in B} \widehat{\mathcal{N}}_n(\hat{\lambda}_n(\beta),\beta). \tag{4.4}$$

As we show in later sections, under H_0 and appropriate conditions,

$$\begin{aligned}&n^{1/2}\widehat{QLR}_n \to_d N(0,\omega_\mu^2),\\ &\quad\text{where } \omega_\mu^2 = E_\mu\big[\exp\big(\gamma_\mu^*(\theta^*)' m(X_i,\theta^*)\big)\\ &\quad - \exp\big(\lambda_\mu^*(\beta^*)' g(X_i,\beta^*)\big)\big]^2,\end{aligned} \tag{4.5}$$

with $\theta^* \in \Theta_\mu^*$ and $\beta^* \in B_\mu^*$.[9]

To form the tests, we also use a variance statistic: $\widehat{\omega}_n^2 = \widehat{\omega}_n^2(\hat{\theta}_n,\hat{\beta}_n)$, where

$$\begin{aligned}\widehat{\omega}_n^2(\theta,\beta) &= n^{-1}\sum_{i=1}^{n}\big[\exp\big(\hat{\gamma}_n(\theta)' m(X_i,\theta)\big) - \exp\big(\hat{\lambda}_n(\beta)' g(X_i,\beta)\big)\big]^2\\ &\quad - \Bigg(n^{-1}\sum_{i=1}^{n}\big[\exp\big(\hat{\gamma}_n(\theta)' m(X_i,\theta)\big)\\ &\quad - \exp\big(\hat{\lambda}_n(\beta)' g(X_i,\beta)\big)\big]\Bigg)^2,\end{aligned} \tag{4.6}$$

and $\hat{\theta}_n$ and $\hat{\beta}_n$ are arbitrary points in $\widehat{\Theta}_n$ and $\widehat{B}_n$, respectively.[10] In practice, different choices of $\hat{\theta}_n$ and $\hat{\beta}_n$ in $\widehat{\Theta}_n$ and $\widehat{B}_n$ typically give the same value for $\widehat{\omega}_n^2(\hat{\theta}_n,\hat{\beta}_n)$ as we find in the Monte Carlo experiments.

Under H_0 and appropriate conditions

$$\widehat{\omega}_n^2 \to_p \omega_\mu^2. \tag{4.7}$$

At this point, it seems that a simple test can be obtained using the studentized QLR statistic: $\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n$ and the standard normal critical value. This is indeed true if we know that ω_μ^2 is bounded away from zero for all relevant data generating processes μ. This is not true if we cannot rule out the μ's for which ω_μ^2 is arbitrarily close or equal to zero. To see why, notice that both $\widehat{QLR}_n$ and $\widehat{\omega}_n^2$ are sample analogue estimators with estimated parameters plugged in. The estimation error in the parameter estimators is dominated by the leading terms in the expansions of $\widehat{QLR}_n$ and $\widehat{\omega}_n^2$ if the leading terms are nondegenerate, that is, if ω_μ^2 is bounded away from zero. But when ω_μ^2 gets arbitrarily close to zero, the estimation error cannot be dominated and will show up in the asymptotic distribution of $\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n$, causing it to be non-normal.

In light of this, we distinguish two testing situations according to whether or not ω_μ^2 is bounded away from zero across all data generating processes μ. The two are specified in Definition NO below. In the definition, we use the variation distance between two probability measures:

$$|P - Q| := \int |dP/dR - dQ/dR| dR, \tag{4.8}$$

where R is any probability measure with respect to which both P and Q are absolutely continuous.[11]

Definition NO. The models $\mathcal{P}$ and $\mathcal{Q}$ are **non-overlapping** if $\inf_{P\in\mathcal{P},Q\in\mathcal{Q}} |P - Q| > 0$ and are **overlapping** otherwise.

Remarks. (a) Our categorization of the model relationships is similar to but different from that in Vuong (1989). We distinguish the two types based on uniform asymptotics—whether $N(0,1)$ can uniformly approximate the finite sample distribution of the studentized quasi-likelihood ratio statistic. Vuong (1989) distinguishes the two types – "strictly nonnested" and "overlapping" – based on pointwise asymptotics. In particular, we treat models $\mathcal{P}$ and $\mathcal{Q}$ as overlapping if it is possible for P_μ^* and Q_μ^* to get arbitrarily close to each other, while Vuong (1989) does not treat them as overlapping as long as $P_\mu^* \neq Q_\mu^*$ under every null distribution μ. Thus our "non-overlapping" concept is stronger than Vuong's (1989) "strictly-nonnested"concept (i.e. $\mathcal{P}\cap\mathcal{Q} = \emptyset$). On the other hand, when both models are variation-closed (that is, closed in the topology defined by the variation metric defined above), being strictly nonnested implies being non-overlapping. A sufficient condition for a moment inequality model $\mathcal{P}$ to be variation-closed is that the moment functions are bounded and continuous in the parameters, as shown in Supplemental Appendix D.

(b) The overlapping case includes the nested case, i.e. $\mathcal{P} \subset \mathcal{Q}$ or $\mathcal{Q} \subset \mathcal{P}$. The results in this paper for overlapping models hold for nested models except for Theorem 2(b).

According to our definition, the two models in (2.3) in Example 1 are non-overlapping if $r_1(X,\theta) \neq r_2(X,\beta)$ for any $\theta \in \Theta$ and $\beta \in B$ and are overlapping otherwise. The two models in Example 2 are overlapping because both models are consistent with a constant f. The models in Example 3 are overlapping because both models are consistent with zero competition effect. The models in Example 4 are overlapping because they are nested. However, it is hard to tell whether or not the models in Example 5 are overlapping or non-overlapping because the moment conditions in the two models have very different structure. It is difficult to know whether the two sets of moment conditions can be simultaneously compatible with one data generating process. In this case, we recommend assuming them to be overlapping to be on the safe side.

[9] By (3.6), ω_μ^2 is invariant to the choice of $\theta^* \in \Theta_\mu^*$ and $\beta^* \in B_\mu^*$.

[10] Notice that the arbitrarily selected points $\hat{\theta}_n$ and $\hat{\beta}_n$ do not necessarily form a random sequence that converges in probability to any points in Θ_μ^* and B_μ^*. This differs from the point selection used in Santos (2011).

[11] We use the variation distance in the Definition NO because a uniform lower bound on ω_μ^2 can be conveniently written in terms of a multiple of the variation distance between the two models. One property of variation distance that leads to such convenience is that it is in variant to the dominating measure, and thus is not tied to a particular μ (Ref. Csiszár, 1975). Another property is that it is a lower bound for the L_2 distance for the densities of P and Q with respect to any μ, and the latter distance forms the main component of ω_μ^2. See the proof of Lemma 4 for details.

4.2. Tests

Let $\alpha \in (0, 1)$. Let $z_{\alpha/2}$ denote the $(1-\alpha/2)$ quantile of the standard normal distribution. We propose tests for non-overlapping models and overlapping models. The test for non-overlapping models does not require a tuning parameter and needs weaker differentiability and moment existence assumptions. However, for the test to have correct asymptotic size, the candidate models should be non-overlapping according to Definition NO. If one applies this test on overlapping models, there may be severe overrejection as our Monte Carlo results show. On the other hand, the test for overlapping models is more general and can be applied to non-overlapping models as well. For easy reference, we name the tests the "non-overlapping test" and the "overlapping test", respectively. Both tests have a one-sided version and a two-sided version, where the two-sided alternative hypothesis is $H_1 : d(\mathcal{P}, \mu) \neq d(\mathcal{Q}, \mu)$ and the one-sided alternative hypothesis is set to be $H_1 : d(\mathcal{P}, \mu) < d(\mathcal{Q}, \mu)$ without loss of generality.

The non-overlapping test. The one-sided version is defined as $\varphi_n^{NO-1}(\alpha) = 1(n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n > z_\alpha)$, and the two-sided version is defined as $\varphi_n^{NO-2}(\alpha) = 1(n^{1/2}|\widehat{QLR}_n|/\widehat{\omega}_n > z_{\alpha/2})$, where α is the nominal size.

The overlapping test. Let b_n be a sequence of positive numbers such that $b_n^{-1} + n^{-1/2}b_n \to 0$. The one-sided version is defined as $\varphi_n^{OL-1}(\alpha) = 1(n^{1/2}\widehat{QLR}_n/(\widehat{\omega}_n \vee n^{-1/2}b_n) > z_\alpha)$, and the two-sided version is defined as $\varphi_n^{OL-2}(\alpha) = 1(n^{1/2}|\widehat{QLR}_n|/(\widehat{\omega}_n \vee n^{-1/2}b_n) > z_{\alpha/2})$, where $a \vee b := \max\{a, b\}$.

It is worthwhile to discuss the intuition behind the asymptotic size control of the two tests. First, the non-overlapping test has correct asymptotic size when applied to non-overlapping models because ω_μ^2 is bounded away from zero for non overlapping models, guaranteeing $n^{1/2}\widehat{QLR}/\widehat{\omega}_n \to_d N(0, 1)$ under H_0. When the models are overlapping, ω_μ^2 is not bounded away from zero, and consequently $n^{1/2}\widehat{QLR}/\widehat{\omega}_n \to_d N(0, 1)$ may or may not hold depending on the unknown data generating process. When it does not hold, the non-overlapping test can overreject. On the other hand, the overlapping tests never overreject asymptotically, thanks to the regularization parameter b_n. Specifically, we show later that $n\widehat{QLR}_n = O_p(1)$ whenever $n^{1/2}\widehat{QLR}/\widehat{\omega}_n \nrightarrow_d N(0, 1)$. This and the fact that b_n is chosen to be a diverging sequence implies that $n^{1/2}\widehat{QLR}_n/(n^{-1/2}b_n) \to_p 0$. As a result, the asymptotic rejection probability of the overlapping test is controlled (below α) even when $n^{1/2}\widehat{QLR}/\widehat{\omega}_n \nrightarrow_d N(0, 1)$.

The regularization parameter is in some sense a critical value for a pretest for $H_{00} : \omega_\mu^2 = 0$. We do not take it to be a finite quantile of the asymptotic distribution of the pretest statistic $n^{1/2}\widehat{\omega}_n$ (under H_{00}) for two reasons. First, the asymptotic distribution of $n^{1/2}\widehat{\omega}_n$ is complicated and difficult to estimate due to both the partial identification and the moment inequalities. Second, a converging critical value in the pretest may not control the asymptotic size of the overall test for H_0. See Shi (forthcoming) for detailed discussions in a related testing problem.

One practical difficulty with the diverging b_n is that there is certain arbitrariness in its choice. The theory in this paper implies that it should satisfy the rate condition $b_n^{-1} + n^{-1/2}b_n \to 0$. However, for a fixed n, this rate condition is not of much help. An optimal finite n choice of b_n should depend on the distributions of the high-order terms in $n\widehat{QLR}_n$ and $n\widehat{\omega}_n^2$. However, in moment inequality models, their distributions are difficult to obtain even asymptotically both due to partial identification, and due to (the unknown slackness of) the inequalities.

Nonetheless, we can borrow some intuition from the point-identified moment *equality* models. Shi (forthcoming) studies such models and the findings therein imply that b_n is needed most when $|(d_\theta - d_m) - (d_\beta - d_g)|$ is large, and least if $(d_\theta - d_m) = (d_\beta - d_g)$. Based on this, we propose the following data-dependent rule-of-thumb choice of b_n:

$$b_n = c \cdot (1 \vee |(d_\theta - \hat{d}_m^b) - (d_\beta - \hat{d}_g^b)|) \cdot \log(\log(n)), \tag{4.9}$$

where $\hat{d}_m^b$ is the number of non-zero components in $\hat{\gamma}_n(\hat{\theta}_n)$ and is used to estimate the number of binding moment conditions for model $\mathcal{P}$, and $\hat{d}_g^b$ is the analogous quantity for model $\mathcal{Q}$. The constant c will be investigated in the Monte Carlo section. Notice that when c is set to zero, the overlapping test reduces to the non-overlapping test.

5. Asymptotic size — non-overlapping case

In this section, we show that the asymptotic size of the non-overlapping test, when applied to non-overlapping models, is correct. To begin, let $\mathcal{H}_0^{no}$ denote the set of null distributions in the case of non-overlapping models. We define $\mathcal{H}_0^{no}$ below. The size of the test for non-overlapping models of nominal size α over $\mathcal{H}_0^{no}$ is

$$SZ_n^{no}(\alpha) = \sup_{\mu \in \mathcal{H}_0^{no}} E_\mu \varphi_n^{NO-j}(\alpha), \tag{5.1}$$

where $j = 1, 2$, recall, stands for the one-sided test and the two-sided test, respectively. We approximate $SZ_n^{no}(\alpha)$ using the asymptotic size:

$$AsySZ^{no}(\alpha) = \limsup_{n\to\infty} SZ_n^{no}(\alpha). \tag{5.2}$$

The following assumption is imposed on the moment functions and is satisfied for all of our examples.

Assumption 4. The moment functions $m(x, \theta)$ and $g(x, \beta)$ are continuously differentiable in θ and β over Θ and B, respectively, for all $x \in \mathcal{X}$.

Let $\mathcal{M}_\mu^* = \max_{\theta\in\Theta} \mathcal{M}_\mu(\gamma_\mu^*(\theta), \theta)$ and $\mathcal{N}_\mu^* = \max_{\beta\in B} \mathcal{N}_\mu(\lambda_\mu^*(\beta), \beta)$. Let $m_i(\theta) = m(X_i, \theta)$ and $g_i(\beta) = g(X_i, \beta)$. For a data distribution μ, and parameters $\theta \in \Theta$ and $\beta \in B$, let

$$S_\mu^m(\gamma, \theta) = E_\mu e^{\gamma' m_i(\theta)} m_i(\theta) m_i(\theta)'$$
$$S_\mu^g(\lambda, \beta) = E_\mu e^{\lambda' g_i(\beta)} g_i(\beta) g_i(\beta)'. \tag{5.3}$$

Let $eig_{\min}(A)$ denote the smallest eigenvalue of a matrix A. For a positive number M, let Γ_M^m denote $N_M(0_{d_m}) \cap (R^{d_p} \times R_+^{d_m-d_p})$, where $N_a(b)$ for a positive scalar a and a d_b-vector b is a closed ball in R^{d_b} centered at b with radius a. Let Γ_M^g denote $N_M(0_{d_g}) \cap (R^{d_q} \times R_+^{d_g-d_q})$. Let $\phi = (\gamma', \theta')'$ and $\psi = (\lambda', \beta')'$. Let "$\wedge$" and "$\vee$" denote the minimum operator and the maximum operator, respectively. Let $N_\varepsilon(\Theta_\mu^*) = \bigcup_{\theta\in\Theta_\mu^*} N_\varepsilon(\theta)$ and $N_\varepsilon(B_\mu^*) = \bigcup_{\beta\in B_\mu^*} N_\varepsilon(\beta)$. We first define the μ space under consideration under both H_0 and H_1 and then define the subset of it for which H_0 holds.

Definition H. The set $\mathcal{H}$ is the set of μ such that

(i) $\{X_i\}_{i=1}^n$ is an i.i.d. sample from μ,

(ii) for all $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ not dependent on μ such that $\sup_{\theta\in\Theta\setminus N_\varepsilon(\Theta_\mu^*)} \mathcal{M}_\mu(\gamma_\mu^*(\theta), \theta) < \mathcal{M}_\mu^* - \delta_\varepsilon$ and $\sup_{\beta\in B\setminus N_\varepsilon(B_\mu^*)} \mathcal{N}_\mu(\lambda_\mu^*(\beta), \beta) < \mathcal{N}_\mu^* - \delta_\varepsilon$,

(iii) $\sup_{\theta\in\Theta} \|\gamma_\mu^*(\theta)\| \vee \sup_{\beta\in B} \|\lambda_\mu^*(\beta)\| \leq M - \delta$,

(iv) $\inf_{\phi\in\Gamma_M^m\times\Theta} eig_{\min}(S_\mu^m(\phi)) \wedge \inf_{\psi\in\Gamma_M^g\times B} eig_{\min}(S_\mu^g(\psi)) > \delta$, and (5.4)

$$
\text{(v) } E_\mu \sup_{\phi \in \Gamma_M^m \times \Theta} \left[e^{(2+\delta)\gamma' m_i(\theta)} + \left\| \frac{\partial e^{\gamma' m_i(\theta)}}{\partial \phi} \right\|^{2+\delta} \right.
$$
$$
\left. + \left\| \frac{\partial^2 e^{\gamma' m_i(\theta)}}{\partial \gamma \partial \phi} \right\|^{2+\delta} + \sum_{j=1}^{d_m} \left\| \frac{\partial^3 e^{\gamma' m_i(\theta)}}{\partial \gamma_j \partial \gamma \partial \phi'} \right\| \right]
$$
$$
+ E_\mu \sup_{\psi \in \Gamma_M^g \times B} \left[e^{(2+\delta)\lambda' g_i(\beta)} + \left\| \frac{\partial e^{\lambda' g_i(\beta)}}{\partial \psi} \right\|^{2+\delta} \right.
$$
$$
\left. + \left\| \frac{\partial^2 e^{\lambda' g_i(\beta)}}{\partial \lambda \partial \psi'} \right\|^{2+\delta} + \sum_{j=1}^{d_g} \left\| \frac{\partial^3 e^{\lambda' g_i(\beta)}}{\partial \lambda_j \partial \lambda \partial \psi'} \right\| \right] < M,
$$

where M and δ are positive constants. The set $\mathcal{H}_0^{no}$ depends on those constants, but for notational simplicity, we suppress their dependence.

Definition H0NO. The set $\mathcal{H}_0^{no}$ is the set of $\mu \in \mathcal{H}$ such that

(i) $d(\mathcal{P}, \mu), d(\mathcal{Q}, \mu) \leq M_1$, for a constant M_1 that does not depend on μ,

(ii) $d(\mathcal{P}, \mu) = d(\mathcal{Q}, \mu)$, and

(iii) μ satisfies Assumption 3.

Remarks. (a) Condition (iii) of Definition H and condition (i) of Definition H0NO are uniform versions of Assumption 1(b), the verification of which is discussed in Section 3.1. Condition (ii) of Definition H rules out weak identification and is standard in the model selection test literature. Condition (iv) of Definition H is the uniform version of Assumption 1(a). Condition (v) of Definition H imposes moment restrictions. The exponential moment restrictions may exclude some interesting cases in practice, but are satisfied in many other cases. For example, they are satisfied by models in Examples 3–5 and by models in Examples 1 and 2 if the variables do not have heavy tails.

(b) Assumption 4 implies Assumption 2(b). Conditions (iii)-(v) of Definition H imply Assumption 1. Therefore, the duality results in Lemma 2 hold for $\mu \in \mathcal{H}$ under Assumptions 2(a) and 4.

In order to derive the asymptotic size of the test, we show the consistency of the set estimators $\widehat{\Theta}_n$ and $\widehat{B}_n$ first. Lemma 3 establishes the consistency of $\widehat{\Theta}_n$ and $\widehat{B}_n$ w.r.t. the *left* Hausdorff distance. The *left* Hausdorff distance between two subsets, A_1, A_2, of a Euclidean space is the maximum distance of any point in A_1 to A_2:

$$
\rho_{lh}(A_1, A_2) = \sup_{a \in A_1} \inf_{a' \in A_2} \|a - a'\|. \tag{5.5}
$$

We call it the *left* Hausdorff distance because its symmetrized version is the Hausdorff distance: $\rho_h(A_1, A_2) = \rho_{lh}(A_1, A_2) \vee \rho_{lh}(A_2, A_1)$. Also define $\rho_{lh}(a, A_2) = \rho_{lh}(\{a\}, A_2)$ for a vector a.

Lemma 3. *Suppose Assumptions 2(a) and 4 hold. Then, under all sequences $\{\mu_n\}_{n=1}^\infty$ such that each $\mu_n \in \mathcal{H}$, we have $\rho_{lh}(\widehat{\Theta}_n, \Theta_{\mu_n}^*) + \rho_{lh}(\widehat{B}_n, B_{\mu_n}^*) \to_p 0$.*

Remark. Lemma 3 shows that all points in $\widehat{\Theta}_n$ approach $\Theta_{\mu_n}^*$. It does not imply that the neighborhoods of all points in $\Theta_{\mu_n}^*$ are visited by $\widehat{\Theta}_n$ eventually. Thus, $\widehat{\Theta}_n$ is not necessarily consistent w.r.t. the standard Hausdorff distance. Consistency w.r.t. ρ_{lh} is sufficient for our purpose.

The following lemma guarantees that the asymptotic variance of $n^{1/2}\widehat{QLR}_n$ is bounded away from zero with non-overlapping models.

Lemma 4. *If the models $\mathcal{P}$ and $\mathcal{Q}$ are non-overlapping, then $\underline{\omega}^2 := \inf_{\mu \in \mathcal{H}_0^{no}} \omega_\mu^2 > 0$.*

The following theorem describes the asymptotic distribution of $n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n$ and shows that the asymptotic size of the test for non-overlapping models is correct.

Theorem 1. *Suppose Assumptions 2(a) and 4 hold and the models are non-overlapping. Then,*

(a) under all sequences $\{\mu_n \in \mathcal{H}_0^{no}\}_{n=1}^\infty$, we have $n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n \to_d N(0, 1)$, and

(b) for $\alpha \in (0, 1)$, $AsySZ^{no}(\alpha) = \alpha$.

6. Asymptotic size — overlapping case

Let $\mathcal{H}_0^{ol}$ denote the set of null distributions in the case of overlapping models. We define $\mathcal{H}_0^{ol}$ below. The size of the test for overlapping models of nominal size α over $\mathcal{H}_0^{ol}$, is

$$
SZ_n^{ol}(\alpha) = \sup_{\mu \in \mathcal{H}_0^{ol}} E_\mu \varphi_n^{OL-j}(\alpha), \tag{6.1}
$$

where, recall, $j = 1, 2$ indicates "one-sided" and "two-sided" respectively. We approximate it using the asymptotic size:

$$
AsySZ^{ol}(\alpha) = \limsup_{n \to \infty} SZ_n^{ol}(\alpha). \tag{6.2}
$$

In the definition of the asymptotic size, the limsup is taken after the $\sup_{\mu \in \mathcal{H}_0^{ol}}$. Thus, in order to obtain $AsySZ^{ol}(\alpha)$, we need to approximate the distribution of the test statistics uniformly well over $\mathcal{H}_0^{ol}$. This is harder to achieve with overlapping models because the asymptotic distributions of $n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n$ and $n\widehat{\omega}_n^2$ under H_0 are discontinuous in ω_μ^2, as discussed in Section 4.1. We seek to approximate the finite sample distributions of the test statistics at all values of ω_μ^2 by deriving the asymptotic distributions under drifting sequences of null distributions $\{\mu_n\}_{n=1}^\infty$. In particular, $n\omega_{\mu_n}^2$ can drift to a finite number or infinity, each case approximating the finite sample situation where ω_μ^2 is close or equal to zero, or ω_μ^2 is bounded away from zero. The idea of using drifting sequences is adopted from Andrews and Guggenberger (2009).

A stronger assumption on the smoothness of the moment functions than Assumption 4 is needed:

Assumption 5. The moment functions $m(x, \theta)$ and $g(x, \beta)$ are three times continuously differentiable in θ and β over Θ and B, respectively, for all $x \in \mathcal{X}$.

Let $\Lambda_{\mu,i}^* = e^{\gamma_\mu^*(\theta^*)' m_i(\theta^*)} - e^{\lambda_\mu^*(\beta^*)' g_i(\beta^*)}$ for arbitrary $\theta^* \in \Theta_\mu^*$ and $\beta^* \in B_\mu^*$. Now we define $\mathcal{H}_0^{ol}$.

Definition H0OL. The set $\mathcal{H}_0^{ol}$ is the set of $\mu \in \mathcal{H}$ such that

(i) $d(\mathcal{P}, \mu) = d(\mathcal{Q}, \mu)$,

(ii) μ satisfies Assumption 3,

(iii) $E_\mu(\omega_\mu^{-1} \Lambda_{\mu,i}^*)^{2+\delta} < M$ if $\omega_\mu^2 > 0$,

(iv) $\mathcal{M}_\mu^* - \mathcal{M}_\mu(\gamma_\mu^*(\theta), \theta) > C \cdot (\rho_{lh}^2(\theta, \Theta_\mu^*) \wedge \delta)$,

$\mathcal{N}_\mu^* - \mathcal{N}_\mu(\lambda_\mu^*(\beta), \beta) > C \cdot (\rho_{lh}^2(\beta, B_\mu^*) \wedge \delta)$, and

$$
\text{(v) } E_\mu \sup_{\phi \in \Gamma_M^m \times \Theta} \left[e^{(2+\delta)\gamma' m_i(\theta)} + \left\| \frac{\partial e^{\gamma' m_i(\theta)}}{\partial \phi} \right\|^{2+\delta} \right.
$$
$$
\left. + \left\| \frac{\partial^2 e^{\gamma' m_i(\theta)}}{\partial \phi \partial \phi} \right\|^{1+\delta} + \sum_{j=1}^{d_m + d_\theta} \left\| \frac{\partial^3 e^{\gamma' m_i(\theta)}}{\partial \phi_j \partial \phi \partial \phi'} \right\| \right]
$$

$$+E_\mu \sup_{\psi\in\Gamma_M^g\times B}\left[e^{(2+\delta)\lambda' g_i(\beta)}+\left\|\frac{\partial e^{\lambda' g_i(\beta)}}{\partial\psi}\right\|^{2+\delta}\right.$$
$$\left.+\left\|\frac{\partial^2 e^{\lambda' g_i(\beta)}}{\partial\psi\partial\psi'}\right\|^{1+\delta}+\sum_{j=1}^{d_g+d_\beta}\left\|\frac{\partial^3 e^{\lambda' g_i(\beta)}}{\partial\psi_j\partial\psi\partial\psi'}\right\|\right]<M, \tag{6.3}$$

where M, C and δ are positive constants. The set $\mathcal{H}_0^{ol}$ depends on M, C and δ, but for notational simplicity, we suppress these arguments.

Remarks. (a) Condition (iv) of Definition H0OL strengthens condition (ii) of Definition H. Such a condition is standard in the model selection literature for point identified models, and is similar to the quadratic minorant condition used in Chernozhukov et al. (2007). It gives us the $n^{-1/2}$-consistency of the set estimators. Condition (v) of Definition H0OL strengthens condition (v) of Definition H, and is usually verified by inspecting the differentiability of the moment functions and the moment existence of the relevant functions of the data.

(b) Condition (iii) of Definition of H0OL helps to characterize the asymptotic behavior of the studentized quasi-likelihood ratio statistic when the standard deviation of $\widehat{QLR}_n$ converges to zero in probability. It is not restrictive because when ω_μ^2 is small, $\Lambda_{\mu,i}^*$ is typically small.

The following Lemma derives the convergence rate of the set estimators under drifting sequences of distributions. The lemma is obtained using the quadratic bounding approach described in the introduction. This approach takes into account the non-differentiability of the population and the sample criterion functions.

Lemma 5. *Suppose Assumptions 2*(a) *and 5 hold. Then, under any drifting sequence $\{\mu_n\in\mathcal{H}\}_{n=1}^\infty$ such that conditions*(iv)–(v) *of Definition H0OL are satisfied, we have $\rho_{lh}(\widehat{\Theta}_n,\Theta_n^*)+\rho_{lh}(\widehat{B}_n,B_n^*)=O_p(n^{-1/2})$.*

Let ω_n^2 abbreviate $\omega_{\mu_n}^2$. We define the drifting sequences of μ's under which the asymptotic behavior of the QLR and variance statistics are studied below. These are the important sequences that determine the asymptotic size of the test.

Definition SEQ. For $\sigma\in[0,\infty]$, let Seq_σ be the set of sequences $\{\mu_{u_n}\in\mathcal{H}_0^{ol}\}_{n=1}^\infty$, such that $\{u_n\}_{n=1}^\infty$ is a subsequence of $\{n\}$, and

$$u_n\omega_{u_n}^2\to\sigma^2. \tag{6.4}$$

Let $Seq=\bigcup_{\sigma\in[0,\infty]}Seq_\sigma$. Notice that we allow σ to take values in the extended real space.

Lemma 6 establishes the asymptotic distributions of the test statistics under drifting sequences in Seq. Part (a) of the lemma includes the completely degenerate case that $\omega_n=0$ for all n and is analogous to Theorem 3.3(i) of Vuong (1989), while part (b) of the lemma includes the nondegenerate case that $\omega_n=\omega$ for some $\omega>0$ for all n and is analogous to Theorem 3.3(ii) of Vuong (1989).

Lemma 6. *Suppose Assumptions 2*(a) *and 5 hold. Then for $\sigma\in[0,\infty]$ and any subsequence $\{u_n\}_{n=1}^\infty$ of $\{n\}$, under any drifting sequence $\{\mu_{u_n}\}_{n=1}^\infty\in Seq_\sigma$,*

(a) *if $\sigma\in[0,\infty)$, $u_n\widehat{\omega}_{u_n}^2=O_p(1)$ and $u_n\widehat{QLR}_{u_n}=O_p(1)$, and*

(b) *if $\sigma=\infty$, $u_n^{1/2}\widehat{QLR}_{u_n}/\omega_{u_n}\to_d N(0,1)$ and $\widehat{\omega}_{u_n}^2/\omega_{u_n}^2\to_p 1$.*

It follows easily from Lemma 6 that $AsySZ^{ol}(\alpha)\leq\alpha$. An extra condition is needed for the test not to be asymptotically conservative and is stated as Assumption 6. Assumption 6 requires the existence of at least one $\mu\in\mathcal{H}_0^{ol}$ under which the pseudo-true distributions from the two models are not the same. Assumption 6 is not restrictive for nonnested models because for a $\mu\in\mathcal{H}_0^{ol}$ that belongs to neither $\mathcal{P}$ or $\mathcal{Q}$, the pseudo-true distributions typically are different except in some pathological cases. Assumption 6 is violated when $\mathcal{P}$ and $\mathcal{Q}$ are nested.

Assumption 6. There exists $\mu\in\mathcal{H}_0^{ol}$, such that $P_\mu^*\neq Q_\mu^*$.

Theorem 2 summarizes the null properties for our test for overlapping models.

Theorem 2. *Suppose Assumptions 2*(a) *and 5 holds. Then, for all $\alpha\in(0,1)$,*

(a) *$AsySZ^{ol}(\alpha)\leq\alpha$, and*

(b) *if Assumption 6 also holds, then $AsySZ^{ol}(\alpha)=\alpha$.*

7. Power properties of the tests

We now show that our model selection tests are consistent against general fixed alternatives and local alternatives that converge to the null at a rate arbitrarily close to $n^{-1/2}$. The results apply to both the overlapping test and the non-overlapping test.

First, we show that our test is consistent against all fixed alternatives under which $d(\mathcal{P},\mu)\neq d(\mathcal{Q},\mu)$. That is, for any $\mu\in\mathcal{H}$ such that $d(\mathcal{P},\mu)<d(\mathcal{Q},\mu)$, the test rejects H_0 in favor of model $\mathcal{P}$ with probability approaching one.

Theorem 3. *Suppose Assumptions 2*(a) *and 4 hold. Then for any $\mu\in\mathcal{H}$ such that $d(\mathcal{P},\mu)<d(\mathcal{Q},\mu)$,*

(a) *$\lim_{n\to\infty}\Pr_\mu\left(n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n>z_{\alpha/2}\right)=1$, and*

(b) *$\lim_{n\to\infty}\Pr_\mu\left(n^{1/2}\widehat{QLR}_n/(\widehat{\omega}_n\vee n^{-1/2}b_n)>z_{\alpha/2}\right)=1$.*

Next, we show that our test is consistent against drifting sequences of alternatives under which $\sqrt{n}(d(\mathcal{P},\mu_n)-d(\mathcal{Q},\mu_n))$ diverges to infinity.

Theorem 4. *Suppose Assumptions 2*(a) *and 4 hold. Then for any sequence $\{\mu_n\in\mathcal{H}\}$ such that μ_n converges weakly to a μ_0 such that $d(\mathcal{P},\mu_0)=d(\mathcal{Q},\mu_0)<\infty$. Suppose also $d(\mathcal{P},\mu_n)\to d(\mathcal{P},\mu_0)$, $d(\mathcal{Q},\mu_n)\to d(\mathcal{Q},\mu_0)$ and $\sqrt{n}(d(\mathcal{P},\mu_n)-d(\mathcal{Q},\mu_n))\to-\infty$; then,*

(a) *$\lim_{n\to\infty}\Pr_{\mu_n}\left(n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n>z_{\alpha/2}\right)=1$, and*

(b) *$\lim_{n\to\infty}\Pr_{\mu_n}\left(n^{1/2}\widehat{QLR}_n/(\widehat{\omega}_n\vee n^{-1/2}b_n)>z_{\alpha/2}\right)=1$.*

8. Simulation

This section reports Monte Carlo results for the missing data example. In this exercise, we investigate (a) the finite sample performance of our tests, (b) the sensitivity of the overlapping test to the tuning parameter c in the data-dependent formula of b_n in (4.9), and (c) the sensitivity of $\widehat{\omega}_n^2(\hat{\theta}_n,\hat{\beta}_n)$ to the choice of $\hat{\theta}_n$ and $\hat{\beta}_n$ in $\widehat{\Theta}_n$ and $\widehat{B}_n$. In the Supplemental Appendix we report additional Monte Carlo results to compare the performance of the overlapping test with the standard χ^2-based test for nested point identified models moment equality models.

The missing data example is a special case of Example 1. Let Y_i be a binary variable that is observed if a selection variable $D_i=1$ and is missing if $D_i=0$. The parameter of interest is $\theta=EY_i$. Let $X_{1,i}$ and $X_{2,i}$ be two candidate instrumental variables, both taking a finite number of values. Let $\bar{Y}_i=Y_iD_i+(1-D_i)$ and $\underline{Y}_i=Y_iD_i$. Then by definition $Y_i\in[\underline{Y}_i,\bar{Y}_i]$. We consider two model

Table 1
Rejection probability in the symmetric nonnested case ($\alpha = 10\%$).

$c\backslash(a_1, a_2)$	(0, 0)	(0.2, 0.2)	(0.5, 0.5)	(0.5, 0)	(0.5, 0.8)
	$n = 250$				
0	(.000, .000)	(.008, .009)	(.037, .039)	(.000, .505)	(.398, .001)
0.4	(.000, .000)	(.008, .009)	(.037, .039)	(.000, .505)	(.398, .001)
	$n = 500$				
0	(.000, .000)	(.016, .015)	(.044, .047)	(.000, .921)	(.658, .000)
0.4	(.000, .000)	(.016, .015)	(.044, .047)	(.000, .921)	(.658, .000)
	$n = 1000$				
0	(.000, .000)	(.033, .028)	(.049, .048)	(.000, 1.000)	(.895, .000)
0.4	(.000, .000)	(.033, .028)	(.049, .048)	(.000, 1.000)	(.895, .000)

The two probabilities in each pair of parentheses are the probability of rejecting H_0 in favor of $\mathcal{P}_1$ and that of rejecting H_0 in favor of $\mathcal{P}_2$, respectively.

comparison problems. In the first problem, the models compared are *nonnested* (but overlapping).[12] That is, for $j = 1, 2$

$$\mathcal{P}_j = \{P : E_P(\theta - \underline{Y}_i | X_{j,i}) \geq 0 \,\&\, E_P(\bar{Y}_i - \theta | X_{j,i}) \geq 0\}. \tag{8.1}$$

In the second problem, the models compared are *nested*, in particular, $\mathcal{P}_2 \subseteq \mathcal{P}_1$:

$$\begin{aligned} \mathcal{P}_1 &= \{P : E_P(\theta - \underline{Y}_i | X_{1,i}) \geq 0 \,\&\, E_P(\bar{Y}_i - \theta | X_{1,i}) \geq 0\}, \\ \mathcal{P}_2 &= \{P : E_P(\theta - \underline{Y}_i | X_{1,i}, X_{2,i}) \geq 0 \\ &\quad \& \, E_P(\bar{Y}_i - \theta | X_{1,i}, X_{2,i}) \geq 0\}. \end{aligned} \tag{8.2}$$

For both problems, we consider the general data generating process (DGP):

$$\begin{aligned} Y_i &= 1\{1 + 1.5^{1/2} \left(a_1 X_{1,i} + a_2 X_{2,i}\right) + u_i \geq 0\} \\ D_i &= 1\{1.5 + 0.5 \left(X_{1,i} + X_{2,i}\right) + v_i \geq 0\}, \quad (u_i, v_i) \sim N(0, I), \end{aligned} \tag{8.3}$$

where $X_{1,i}$ and $X_{2,i}$ follow independent multinomial distributions. The parameters a_1 and a_2 measure how endogenous the two instruments are and thus indicate how far each model is from the DGP. Both the nonnested and the nested problems fit in the framework of Example 1 because $\mathcal{P}_j$ can be written as

$$\begin{aligned} \mathcal{P}_j = \{P : E_P[(\theta - \underline{Y}_i) 1(Z_{j,i} = z)] \geq 0 \,\&\, E_P[(\bar{Y}_i - \theta) 1(Z_{j,i} = z)] \\ \geq 0 \,\forall z \in \mathcal{Z}_j\}, \end{aligned} \tag{8.4}$$

where $Z_{j,i}$ is the conditioning variable/vector of model j and $\mathcal{Z}_j$ is the known discrete support of $Z_{j,i}$. For the nonnested problem, we consider the test of H_0 against the two-sided alternative, while for the nested problem, we test H_0 against the one-sided alternative $H_1 : d(\mathcal{P}_1, \mu) < d(\mathcal{P}_2, \mu)$.

For the nonnested problem, we consider two designs of $(X_{1,i}, X_{2,i})$, a symmetric one and an asymmetric one. In the symmetric design, $X_{1,i}$ and $X_{2,i}$ both follow a multinomial distribution that puts equal probability on the points in $\{0, 1\}$. The symmetry of the two models makes it easy to specify the null DGP's: to impose H_0, we can simply set $a_1 = a_2$. Varying the magnitude of a_1 $(=a_2)$ allows us to vary the magnitude of ω_μ^2. In particular, the larger $a_1 (=a_2)$ is, the further away ω_μ^2 is from zero. The alternative DGP's are also easy to specify: when $a_1 > a_2 \geq 0$, $\mathcal{P}_2$ is better (less misspecified) than $\mathcal{P}_1$, and vice versa. To cover a variety of cases, we consider five pairs of (a_1, a_2): $(0, 0)$, $(0.2, 0.2)$, $(0.5, 0.5)$, $(0.5, 0)$, $(0.5, 0.8)$, three different sample sizes: $n = 250, 500, 1000$ and two different choices of tuning parameter $c = 0$ and 0.4. Note that for $c = 0$, the test is the non-overlapping test and for $c > 0$, it is the overlapping one. The number of simulation repetitions is 5000. The nominal size of the tests is 10%.

Table 1 shows the rejection probabilities for the symmetric nonnested design. The first three columns show the rejection probabilities under the null. Ideally, under H_0, the probability of rejecting H_0 in favor of either model should be at or below 5%. As we can see from the first three columns, this requirement is satisfied, indicating that our test controls size well in finite samples. The last two columns show the rejection probability under the alternative. For the fourth column, $\mathcal{P}_2$ is better and for the last column, $\mathcal{P}_1$ is better. As we can see, our test selects the better model with nontrivial probability while rarely selects the worse model. Also the probability of rejecting H_0 in favor of the better model increases with the sample size as expected from the power results. In addition, varying the tuning parameter c in the range that we consider has no effect on the rejection probabilities. The robustness to c is a result of the symmetry of the two models compared, and unfortunately is not a generic feature of our test, as shown in the next design.

Now we consider the asymmetric nonnested design, where $X_{1,i}$ has the same distribution as above, but $X_{2,i}$ follows a multinomial distribution that puts equal probability on J equally spaced points in the interval $[-1, 1]$ (including the end points). In this case, setting $a_1 = a_2$ does not guarantee that H_0 hold due to the asymmetry of the two models. However, we can still ensure H_0 by setting $a_1 = a_2 = 0$, and ensure that H_0 does not hold by setting $a_1 \neq 0$ and $a_2 = 0$. The parameter J controls the degree of the asymmetry. We report Monte Carlo results for two values of (a_1, a_2): $(0, 0)$ and $(0.5, 0)$, three values of J: 3, 7 and 11 and four c values: 0, 0.2, 0.25 and 0.3. The sample sizes and number of simulation repetitions are the same as the previous design.

Table 2 shows the rejection probabilities for the asymmetric nonnested design. The first three columns show the rejection probabilities under the null and the last three columns show those under the alternative. Comparing to the previous table, we first observe that the over-lapping test ($c = 0$) has over-rejection for the most asymmetric design ($J = 11$) at all three sample sizes. For the non-overlapping test ($c > 0$), the rejection probabilities are somewhat sensitive to c both under the null and under the alternative in the most asymmetric design, but not so much in the less asymmetric designs. Overall, Table 2 shows that the overlapping test with $c = 0.25$ and $c = 0.3$ has decent performance.

Lastly, we consider the nested problem in (8.2). We let the distribution of X_{1i} and X_{2i} be the same as the asymmetric nonnested design above. We consider two values of (a_1, a_2): $(0, 0)$ and $(0, 0.25)$, each representing the null and the alternative respectively. The same J values and c values as above are considered. Note that for the same J, the two nested models in (8.2) are much more asymmetric than the two nonnested models in (8.1) because the model 2 in the nested case involves $2J$ rather than J unconditional moment restrictions while the model 1 still only contains 2 unconditional moment restrictions. Because our data-dependent choice

[12] Even though without additional information the two models are overlapping, they become non-overlapping if we add the maintained assumption that $\min\{cov(X_{1,i}, Y_i), cov(X_{2,i}, Y_i)\} > \eta$ for some $\eta > 0$. Adding this maintained assumption does not affect how our tests should be implemented. Thus, the nonnested results below when this maintained assumption are satisfied (i.e. when $a_1, a_2 > 0$) also demonstrate how the tests perform in a non-overlapping testing scenario.

Table 2
Rejection probability in the asymmetric nonnested case ($\alpha = 10\%$).

c	$(a_1, a_2) = (0, 0)$			$(a_1, a_2) = (0.5, 0)$		
	$J = 3$	$J = 7$	$J = 11$	$J = 3$	$J = 7$	$J = 11$
	$n = 250$					
0	(.002, .000)	(.044, .000)	(.226, .000)	(.000, .346)	(.001, .138)	(.009, .047)
0.2	(.002, .000)	(.044, .000)	(.188, .000)	(.000, .346)	(.001, .138)	(.007, .046)
0.25	(.002, .000)	(.043, .000)	(.124, .000)	(.000, .346)	(.001, .138)	(.005, .041)
0.3	(.002, .000)	(.041, .000)	(.069, .000)	(.000, .346)	(.001, .137)	(.004, .029)
	$n = 500$					
0	(.001, .000)	(.031, .000)	(.194, .000)	(.000, .815)	(.000, .559)	(.000, .347)
0.2	(.001, .000)	(.031, .000)	(.166, .000)	(.000, .815)	(.000, .559)	(.000, .347)
0.25	(.001, .000)	(.031, .000)	(.102, .000)	(.000, .815)	(.000, .559)	(.000, .337)
0.3	(.001, .000)	(.029, .000)	(.049, .000)	(.000, .815)	(.000, .559)	(.000, .301)
	$n = 1000$					
0	(.002, .000)	(.028, .000)	(.165, .000)	(.000, .996)	(.000, .973)	(.000, .910)
0.2	(.002, .000)	(.028, .000)	(.139, .000)	(.000, .996)	(.000, .973)	(.000, .910)
0.25	(.002, .000)	(.027, .000)	(.078, .000)	(.000, .996)	(.000, .973)	(.000, .910)
0.3	(.002, .000)	(.025, .000)	(.039, .000)	(.000, .996)	(.000, .973)	(.000, .905)

The two probabilities in each pair of parentheses are the probability of rejecting H_0 in favor of $\mathcal{P}_1$ and that of rejecting H_0 in favor of $\mathcal{P}_2$, respectively.

Table 3
Rejection probability of the one-sided tests in the nested case ($\alpha = 5\%$).

c	$(a_1, a_2) = (0, 0)$			$(a_1, a_2) = (0, 0.25)$		
	$J = 3$	$J = 7$	$J = 11$	$J = 3$	$J = 7$	$J = 11$
	$n = 250$					
0	.025	.523	.948	.477	.826	.987
0.2	.025	.249	.274	.477	.557	.484
0.25	.025	.116	.113	.476	.360	.259
0.3	.024	.049	.040	.467	.216	.129
	$n = 500$					
0	.020	.465	.924	.825	.936	.994
0.2	.020	.204	.186	.825	.752	.616
0.25	.020	.086	.056	.824	.567	.349
0.3	.020	.032	.013	.818	.386	.160
	$n = 1000$					
0	.017	.400	.886	.994	.994	1.000
0.2	.017	.163	.137	.994	.947	.844
0.25	.017	.062	.031	.994	.854	.627
0.3	.017	.023	.006	.993	.718	.383

The probabilities are the probability of the one-sided tests rejecting H_0 in favor of $\mathcal{P}_1$.

of b_n is adaptive to the asymmetry, we shall see that the large asymmetry does not have much ill-effect on the size property of our test.

Table 3 shows the results. Since the models are nested ($\mathcal{P}_2 \subseteq \mathcal{P}_1$), only the one-sided alternative $H_1 : d(\mathcal{P}_1, \mu) > d(\mathcal{P}_2, \mu)$ is of interest. Thus, the table reports the rejection probabilities for the one-sided tests. As we can see, the pattern is similar to the nonnested asymmetric design. Both the non-overlapping test and the overlapping test have good size and power in the mildly asymmetric design ($J = 3$). The non-overlapping test ($c = 0$) starts to over-reject as the asymmetry increases. Similar behavior is also observed for the overlapping test with small c ($c = 0.2$). We find that $c = 0.25$ has acceptable performance at all sample sizes across all J's considered.

To sum up, the Monte Carlo shows that (a) both the overlapping test and the non-overlapping test have good finite sample size and power properties and the performance for the non-overlapping test is not sensitive to c, when the two models compared have similar dimensions; and (b) the overlapping test and the non-overlapping test with small c over-reject when the two models are very different in their numbers of restrictions. In the latter cases, we recommend $c = 0.25$.

In the Monte Carlo exercises above, to find $\hat{\theta}_n$ and $\hat{\beta}_n$, we use the *fminbnd* function in Matlab. The *fminbnd* function takes an upper and a lower bound for the parameter. When the bounds are set differently, the function can converge to different minimizers of the criterion function when the minimizer is not unique. That allows us to investigate the sensitivity of our test to the choice of $\hat{\theta}_n$ and $\hat{\beta}_n$ by comparing $\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n$ computed using two different sets of bounds in *fminbnd*. We find that $\hat{\theta}_n$ (or $\hat{\beta}_n$) can be sensitive to the bounds when the model is correctly specified, i.e., when $a_1 = 0$ (or $a_2 = 0$) but not sensitive otherwise. Even when $\hat{\theta}_n$ is sensitive, we find $\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n$ barely differs across the two sets of bounds. For example, in the nested design with $(\alpha_1, \alpha_2) = (0, 0)$, $J = 11$ and $n = 1000$, the frequency that $\hat{\theta}_n$ from the two sets of bounds differ by more than 0.001 is 32%, while the frequency that $\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n$ differ by more than 0.0001 is 0%. This confirms that we do not need to compute all the maximizers to implement the test.

The computational cost of the test is relatively low. In the simulation example described above, it takes around one second to run one simulation iteration on a regular desktop. The speed does not increase with the sample size in the range that we considered. Of course, for models with covariates and more parameters, computation time can be longer, but we expect it to be in a reasonable range for the reasons discussed in the introduction.

Acknowledgments

The author is deeply indebted to Donald Andrews for guidance and encouragement. The author benefited most from the comments of Quang H. Vuong, Xiaohong Chen, Edward Vytlacil, Yuichi Kitamura, Taisuke Otsu, and Peter Phillips. The author also appreciates the helpful comments from three anonymous referees and the anonymous associate editor, as well as those from seminar participants in all the schools where this paper has been presented as a job market paper. The financial support from a Carl Arvid Anderson Fellowship is gratefully acknowledged. The revision of the paper was partly funded by the Wisconsin Alumni Research Foundation via the Graduate School Fall Competition Grant.

Appendix

Throughout the appendix, we replace μ_n with n when μ_n is in a subscript and it does not cause confusion to do so. For example, we write $\gamma^*_{\mu_n}(\theta)$ as $\gamma^*_n(\theta)$. Let $\hat{\phi}_n(\theta) = (\hat{\gamma}_n(\theta)', \theta')'$ and $\phi^*_n(\theta) = (\gamma^*_n(\theta)', \theta')'$, and $\hat{\psi}_n(\beta)$ and $\psi^*_n(\beta)$ be defined analogously. We let "r.h.s." denote "right-hand-side" and "l.h.s." denote "left-hand-side".

Let "LLN" denote the weak law of large number for row-wise i.i.d. triangular arrays. The weak law of large number we use here is Theorem 2 in Andrews (1988). Theorem 2 in Andrews (1988) is a law of large numbers for L^1-mixingale triangular arrays. Rowwise i.i.d. triangular arrays are trivially L^1-mixingales. The uniform integrability condition required in that theorem is guaranteed by the moment existence conditions in this paper.

Appendix A. Auxiliary lemmas

We first present a few auxiliary lemmas, the proofs of which are given in Supplemental Appendix G. Lemma A.1 is an instrumental result for the uniform stochastic boundedness of empirical processes, which is useful for establishing Lemmas A.2–A.3. Lemma A.2 establishes the uniform convergence and rate of convergence of various stochastic processes, which is useful for proving the main lemmas and theorems. The proof of Lemma A.2 makes use of the uniform generic convergence results in Andrews (1992) and the empirical process results reviewed in Andrews (1994). Lemma A.3 establishes the uniform consistency of $\hat{\gamma}_n(\theta)$, the rate of convergence of $\hat{\gamma}_n(\theta)$, and the continuity of $\gamma_n^*(\theta)$.

Lemmas A.2–A.4 are stated in terms of $\{n\}$, but because they only require termwise assumptions on the sequence $\{\mu_n\}_{n=1}^{\infty}$, their conclusions hold with $\{n\}$ replaced with any subsequence of $\{n\}$.

Lemma A.1. *Consider the triangular array of empirical processes $\{\nu_n(\phi) : \phi \in \Phi\}_{n=1}^{\infty}$. If* (i) *$(\Phi, \rho)$ is a totally bounded pseudo-metric space,* (ii) *$\nu_n(\phi)$ is stochastically equicontinuous w.r.t. ρ and* (iii) *for every $\phi \in \Phi$, $\|\nu_n(\phi)\| = O_p(1)$, then $\sup_{\phi\in\Phi} \|\nu_n(\phi)\| = O_p(1)$.*[13]

Lemma A.2. *Suppose Assumptions* 2(a) *and* 4 *hold. Under any sequence $\{\mu_n\}_{n=1}^{\infty}$ such that each μ_n satisfies conditions* (i) *and* (iii)–(v) *of Definition* H*, we have*

(a) *the triangular array of empirical processes $\{\nu_n^0(\phi) := n^{1/2}(\widehat{\mathcal{M}}_n(\phi) - \mathcal{M}_{\mu_n}(\phi)) : \phi \in \Gamma_M^m \times \Theta\}$ is stochastically equicontinuous w.r.t. the Euclidean distance,*

(b) $\sup_{\phi\in\Gamma_M^m\times\Theta} |n^{1/2}(\widehat{\mathcal{M}}_n(\phi) - \mathcal{M}_{\mu_n}(\phi))| = O_p(1)$,

(c) *the triangular array of empirical processes $\{\nu_n^1(\phi) := n^{1/2}(\partial\widehat{\mathcal{M}}_n(\phi)/\partial\gamma - \partial\mathcal{M}_{\mu_n}(\phi)/\partial\gamma) : \phi \in \Gamma_M^m \times \Theta\}$ is stochastically equicontinuous w.r.t. the Euclidean distance,*

(d) $\sup_{\phi\in\Gamma_M^m\times\Theta} \|n^{1/2}(\partial\widehat{\mathcal{M}}_n(\phi)/\partial\gamma - \partial\mathcal{M}_{\mu_n}(\phi)/\partial\gamma)\| = O_p(1)$,

(e) *for all random sequences $\{\phi_{1,n} \in \Gamma_M^m \times \Theta\}_{n=1}^{\infty}$ and $\{\phi_{2,n} \in \Gamma_M^m \times \Theta\}_{n=1}^{\infty}$ such that $\|\phi_{1,n} - \phi_{2,n}\| \to_p 0$, we have*

$$\|\partial^2\widehat{\mathcal{M}}_n(\phi_{1,n})/\partial\gamma\partial\gamma' - \partial^2\mathcal{M}_{\mu_n}(\phi_{2,n})/\partial\gamma\partial\gamma'\| \to_p 0$$

$$|\widehat{\mathcal{M}}_n(\phi_{1,n}) - \mathcal{M}_{\mu_n}(\phi_{2,n})| \to_p 0, \quad \textit{and}$$

(f) *parts*(a)–(e) *hold with Θ, γ, ϕ, $\mathcal{M}$ and m replaced with B, λ, ψ, $\mathcal{N}$ and g, respectively.*

Lemma A.3. *Suppose Assumptions* 2(a) *and* 4 *hold. Under any sequence $\{\mu_n\}_{n=1}^{\infty}$ such that each μ_n satisfies conditions* (i) *and* (iii)–(v) *of Definition* H*, we have*

(a) *for any two random sequences $\{\theta_{1,n} \in \Theta\}_{n=1}^{\infty}$ and $\{\theta_{2,n} \in \Theta\}_{n=1}^{\infty}$ such that $\|\theta_{1,n} - \theta_{2,n}\| \to_p 0$,*

$$\|\hat{\gamma}_n(\theta_{1,n}) - \gamma_n^*(\theta_{2,n})\| \to_p 0,$$

(b) $\sup_{\theta\in\Theta} \|\hat{\gamma}_n(\theta) - \gamma_n^*(\theta)\| = O_p(n^{-1/2})$,

(c) *for any two random sequences $\{\theta_{1,n} \in \Theta\}_{n=1}^{\infty}$ and $\{\theta_{2,n} \in \Theta\}_{n=1}^{\infty}$ such that $\|\theta_{1,n} - \theta_{2,n}\| \to_p 0$,*

$$\|\gamma_n^*(\theta_{1,n}) - \gamma_n^*(\theta_{2,n})\| = O_p(\|\theta_{1,n} - \theta_{2,n}\|), \quad \textit{and}$$

(d) *parts* (a)–(c) *hold with θ, Θ, γ, ϕ, $\mathcal{M}$, m replaced with β, B, λ, ψ, $\mathcal{N}$, g.*

Lemma A.4. *Suppose Assumptions* 2(a) *and* 5 *hold. Then, under any sequence $\{\mu_n\}_{n=1}^{\infty}$ such that each μ_n satisfies conditions* (i) *and* (iii)–(v) *of Definition* H *and condition* (v) *of Definition* H0OL,

(a) *for any two random sequences $\{\phi_{1,n} \in \Gamma_M^m \times \Theta\}_{n=1}^{\infty}$ and $\{\phi_{2,n} \in \Gamma_M^m \times \Theta\}_{n=1}^{\infty}$ such that $\|\phi_{1,n} - \phi_{2,n}\| \to_p 0$,*

$$\|\partial^2\widehat{\mathcal{M}}_n(\phi_{1,n})/\partial\phi\partial\phi' - \partial^2\mathcal{M}_{\mu_n}(\phi_{2,n})/\partial\phi\partial\phi'\| \to_p 0, \quad \textit{and}$$

(b) *part* (a) *hold with Θ, ϕ, $\mathcal{M}$ and m replaced with B, ψ, $\mathcal{N}$ and g.*

Appendix B. Proof of the theorems

Proof of Theorem 1. (a) Let $\hat{\theta}_n \in \widehat{\Theta}_n$ and $\hat{\beta}_n \in \widehat{B}_n$ be those that satisfy $\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n) = \widehat{\omega}_n^2$. Then, part (a) is implied by:

$$n^{1/2}\widehat{QLR}_n/\omega_n \to_d N(0, 1), \quad \text{and} \tag{B.1}$$

$$\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n)/\omega_n^2 \to_p 1. \tag{B.2}$$

Next, we show (B.1) and (B.2).

Let $\theta_n^* \in \Theta_n^*$ and $\beta_n^* \in B_n^*$ satisfy $\|\hat{\theta}_n - \theta_n^*\| \le \rho_{lh}(\hat{\theta}_n, \Theta_n^*) + o(1)$ and $\|\hat{\beta}_n - \beta_n^*\| \le \rho_{lh}(\hat{\beta}_n, B_n^*) + o(1)$. Then, Lemmas 3 and A.3(a) imply that

$$\begin{aligned} &\|\hat{\phi}_n(\hat{\theta}_n) - \phi_n^*(\theta_n^*)\| \to_p 0 \quad \text{and} \\ &\|\hat{\psi}_n(\hat{\beta}_n) - \psi_n^*(\beta_n^*)\| \to_p 0. \end{aligned} \tag{B.3}$$

First, we show (B.1). Observe that

$$\begin{aligned} \omega_n^{-1}n^{1/2}\widehat{QLR}_n &= \omega_n^{-1}n^{-1/2}\sum_{i=1}^{n}\big[\exp(\hat{\gamma}_n(\hat{\theta}_n)'m_i(\hat{\theta}_n)) \\ &\quad - \exp(\hat{\lambda}_n(\hat{\beta}_n)'g_i(\hat{\beta}_n))\big] \\ &= \omega_n^{-1}n^{-1/2}\sum_{i=1}^{n}(\Lambda_{n,i}^*) + A_{n,1} + A_{n,2}, \end{aligned} \tag{B.4}$$

where $\Lambda_{n,i}^* = e^{\gamma_n^*(\theta_n^*)'m_i(\theta_n^*)} - e^{\lambda_n^*(\beta_n^*)'g_i(\beta_n^*)}$, $A_{n,1} = \frac{1}{\sqrt{n\omega_n^2}}\sum_{i=1}^n \big[e^{\hat{\gamma}_n(\hat{\theta}_n)'m_i(\hat{\theta}_n)} - e^{\gamma_n^*(\theta_n^*)'m_i(\theta_n^*)}\big]$ and $A_{n,2} = \frac{1}{\sqrt{n\omega_n^2}}\sum_{i=1}^n \big[e^{\lambda_n^*(\beta_n^*)'g_i(\beta_n^*)} - e^{\hat{\lambda}_n(\hat{\beta}_n)'g_i(\hat{\beta}_n)}\big]$.

By the Lyapunov CLT for triangular arrays,

$$\omega_n^{-1}n^{-1/2}\sum_{i=1}^{n}(\Lambda_{n,i}^*) \to_d N(0, 1). \tag{B.5}$$

The CLT applies because (a) $E_n\Lambda_{n,i}^* = 0$ by Definition H0NO and Lemma 1(b), (b) $\omega_n^{-2}E_n(\Lambda_{n,i}^* - E_n\Lambda_{n,i}^*)^2 = 1$ by the definition of ω_n^2 and (c) the Lyapunov condition holds, that is to say, $E_n(\omega_n^{-1}\Lambda_{n,i}^*)^{2+\delta} \le \omega^{-2-\delta}E_n(\Lambda_{n,i}^*)^{2+\delta} < \infty$ by Lemma 4 and condition (v) in (5.4).

It is left to show $A_{n,1} = o_p(1)$ and $A_{n,2} = o_p(1)$ before we can conclude that (B.1) holds. It suffices to show $A_{n,1} = o_p(1)$ since the arguments for $A_{n,2} = o_p(1)$ are analogous. Because we do not have convergence rates for $\widehat{\Theta}_n$ and $\widehat{B}_n$ under the conditions of the current theorem, the usual approach of doing a second-order Taylor expansion of $\exp(\hat{\gamma}_n(\hat{\theta}_n)'m_i(\hat{\theta}_n))$ around $\phi_n^*(\theta_n^*)$ does not go through. Instead, we show $A_{n,1} = o_p(1)$ by bounding $A_{n,1}$ from both above and below by $o_p(1)$. The lower bound of $A_{n,1}$ is obtained by replacing $\hat{\theta}_n$ with θ_n^* in the expression of $A_{n,1}$ and using the convergence rate result for $\hat{\gamma}_n(\cdot)$ (Lemma A.3(b)):

$$A_{n,1} \ge \omega_n^{-1}n^{-1/2}\sum_{i=1}^{n}\big[\exp(\hat{\gamma}_n(\theta_n^*)'m_i(\theta_n^*)) - \exp(\gamma_n^*(\theta_n^*)'m_i(\theta_n^*))\big]$$

[13] Note that here, Φ denotes the space of ϕ. In the main sections of this paper, Φ stands for the c.d.f. of the standard normal distribution. Hopefully, there is no confusion caused by this abuse of notation.

$$\begin{aligned}
&= \omega_n^{-1}\big[\partial\widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))/\partial\gamma'\big]\big[n^{1/2}(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*))\big]\\
&\quad + \omega_n^{-1}n^{1/2}\big(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\big)'\big[\partial^2\widehat{\mathcal{M}}_n(\tilde\phi_n)/\partial\gamma\partial\gamma'\big]\\
&\quad \times \big(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\big)\\
&\geq \omega_n^{-1}\big[\partial\widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))/\partial\gamma' - \partial\mathcal{M}_{\mu_n}(\phi_n^*(\theta_n^*))/\partial\gamma'\big]\\
&\quad \times \big[n^{1/2}(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*))\big]\\
&\quad + \omega_n^{-1}n^{1/2}\big(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\big)'\big[\partial^2\widehat{\mathcal{M}}_n(\tilde\phi_n)/\partial\gamma\partial\gamma'\big]\\
&\quad \times \big(\hat\gamma_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\big) = o_p(1), \qquad \text{(B.6)}
\end{aligned}$$

where $\tilde\phi_n$ lies on the line segment joining $\phi_n^*(\theta_n^*)$ and $\hat\phi_n(\theta_n^*)$, the first inequality holds because $\widehat{\Theta}_n$ is a maximizer of $\widehat{\mathcal{M}}_n(\hat\gamma_n(\cdot), \cdot)$, the first equality holds by a Taylor expansion of $\exp(\hat\gamma_n(\theta_n^*)'m_i(\theta_n^*))$ around $\gamma_n^*(\theta_n^*)$, and the second inequality holds because $\big(\partial\mathcal{M}_{\mu_n}(\phi_n^*(\theta_n*))/\partial\gamma'\big)\,\gamma_n^*(\theta_n^*) = 0$ and $\partial\mathcal{M}_{\mu_n}(\phi_n^*(\theta_n*))/\partial\gamma_j \begin{cases} = 0 & \text{for } j \leq d_p\\ \geq 0 & \text{for } j > d_p\end{cases}$, both being the Kuhn–Tucker conditions from the minimization problem $\min_{\gamma\in R^{d_p}\times R_+^{d_m-d_p}} \mathcal{M}_{\mu_n}(\gamma, \theta_n^*)$, and the second equality holds by Lemmas 4 and A.2(d)–(e), A.3(b) and condition (v) of (5.4).

The upper bound of $A_{n,1}$ is obtained by replacing $\hat\gamma_n$ with γ_n^* in the expression of $A_{n,1}$ and applying Lemma A.2(a):

$$\begin{aligned}
A_{n,1} &\leq \omega_n^{-1}n^{-1/2}\sum_{i=1}^{n}\big[\exp(\gamma_n^*(\hat\theta_n)'m_i(\hat\theta_n)) - \exp(\gamma_n^*(\theta_n^*)'m_i(\theta_n^*))\big]\\
&= \omega_n^{-1}\big[\nu_n^0(\phi_n^*(\hat\theta_n)) + n^{1/2}\big(\mathcal{M}_{\mu_n}(\phi_n^*(\hat\theta_n)) - \mathcal{M}_{\mu_n}(\phi_n^*(\theta_n^*))\big)\\
&\quad - \nu_n^0(\phi_n^*(\theta_n^*))\big]\\
&\leq \omega_n^{-1}[\nu_n^0(\phi_n^*(\hat\theta_n)) - \nu_n^0(\phi_n^*(\theta_n^*))] = o_p(1), \qquad \text{(B.7)}
\end{aligned}$$

where the first inequality holds because $\hat\gamma_n(\hat\theta_n)$ is the minimizer of $\widehat{\mathcal{M}}_n(\cdot, \hat\theta_n)$, the first equality holds by adding and subtracting terms to form the empirical process $\{\nu_n^0(\phi) : \phi \in \Gamma_M^m \times \Theta\}_{n=1}^{\infty}$ (defined in Lemma A.2(a)), the second inequality holds because θ_n^* is a maximizer of $\mathcal{M}_{\mu_n}(\phi_n^*(\theta))$, and the second equality holds by Lemmas 3, 4 and A.2(a) and A.3(c).

Therefore, $A_{n,1} = o_p(1)$.

Next, we show (B.2). By a mean-value expansion of $\exp(\hat\gamma_n(\hat\theta_n)' m_i(\hat\theta_n))$ around $\phi_n^*(\theta_n^*)$, we have

$$\begin{aligned}
\omega_n^{-2}\widehat{\omega}_n^2(\hat\theta_n, \hat\beta_n) &= -\omega_n^{-2}\widehat{QLR}_n^2 + \omega_n^{-2}n^{-1}\sum_{i=1}^{n}(\Lambda_{n,i}^*)^2\\
&\quad + 2\omega_n^{-2}n^{-1}\sum_{i=1}^{n}\left[\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi'}(\hat\phi_n(\hat\theta_n) - \phi_n^*(\theta_n^*))\right.\\
&\quad \left. - \frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi'}(\hat\psi_n(\hat\beta_n) - \psi_n^*(\beta_n^*))\right](\Lambda_{n,i}^*)\\
&\quad + \omega_n^{-2}n^{-1}\sum_{i=1}^{n}\left[\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi'}(\hat\phi_n(\hat\theta_n) - \phi_n^*(\theta_n^*))\right.\\
&\quad \left. - \frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi'}(\hat\psi_n(\hat\beta_n) - \psi_n^*(\beta_n^*))\right]^2\\
&\equiv W_{n,0} + W_{n,1} + W_{n,2} + W_{n,3}, \qquad \text{(B.8)}
\end{aligned}$$

where $(\tilde\gamma_n', \tilde\theta_n')'$ lies on the line segment joining $\hat\phi_n(\hat\theta_n)$ and $\phi_n^*(\theta_n^*)$ and $(\tilde\lambda_n', \tilde\beta_n')'$ lies on the line segment joining $\hat\psi_n(\hat\beta_n)$ and $\psi_n^*(\beta_n^*)$.

The first summand $W_{n,0} \equiv -\omega_n^{-2}\widehat{QLR}_n^2 = o_p(1)$ by (B.1). The second summand $W_{n,1} \equiv \omega_n^{-2}n^{-1}\sum_{i=1}^{n}(\Lambda_{n,i}^*)^2 \to_p 1$ by LLN. The LLN applies because (a) $E_n\omega_n^{-2}(\Lambda_{n,i}^*)^2 = 1$, and (b) $\sup_{n\geq1} E_n (\omega_n^{-1}\Lambda_{n,i}^*)^{2+\delta} \leq \omega^{-2-\delta}\sup_{n\geq1} E_n(\Lambda_{n,i}^*)^{2+\delta} < \infty$, by condition (v) in (5.4) and Lemma 4.

The summand $W_{n,3}$ in the last line of (B.8) is $o_p(1)$ because

$$\begin{aligned}
0 &\leq W_{n,3}\\
&\leq 2\underline{\omega}^{-2}(\hat\phi_n(\hat\theta_n) - \phi_n^*(\theta_n^*))\left(n^{-1}\sum_{i=1}^{n}\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi}\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi'}\right)\\
&\quad \times (\hat\phi_n(\hat\theta_n) - \phi_n^*(\theta_n^*)) + 2\underline{\omega}^{-2}(\hat\psi_n(\hat\beta_n) - \psi_n^*(\beta_n^*))\\
&\quad \times \left(n^{-1}\sum_{i=1}^{n}\frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi}\frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi'}\right)(\hat\psi_n(\hat\beta_n) - \psi_n^*(\beta_n^*))\\
&= o_p(1), \qquad \text{(B.9)}
\end{aligned}$$

where the second inequality holds by the inequality, $(a+b)^2 \leq 2a^2 + 2b^2$ and Lemma 4, and the equality holds by (B.3) and

$$\begin{aligned}
E_n\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi}\frac{\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}}{\partial\phi'}\right\| &\leq E_n\|\partial e^{\tilde\gamma_n' m_i(\tilde\theta_n)}/\partial\phi\|^2 \leq M\\
E_n\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi}\frac{\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}}{\partial\psi'}\right\| &\leq E_n\|\partial e^{\tilde\lambda_n' g_i(\tilde\beta_n)}/\partial\psi'\|^2\\
&\leq M, \qquad \text{(B.10)}
\end{aligned}$$

which holds by the triangular inequality, the equality $\|aa'\| = \|a\|^2$ and condition (v) in (5.4).

The summand $W_{n,2}$ in the last line of (B.8) is $o_p(1)$ because, by the Cauchy–Schwarz inequality, $0 \leq \left|W_{n,2}\right| \leq 2[W_{n,1}\cdot W_{n,3}]^{1/2}$.

Therefore, (B.2) holds.

(b) For this part, we focus on the two-sided test. Arguments for the one-sided test is the same. Let $\{\mu_n \in \mathcal{H}_0^{no}\}_{n=1}^{\infty}$ satisfy $\Pr_n\big(n^{1/2}|\widehat{QLR}_n|/\widehat{\omega}_n > z_{\alpha/2}\big) \geq SZ_n^{no}(\alpha) - o(1)$. Such a sequence always exists. Then,

$$\limsup_{n\to\infty}\Pr_n\big(n^{1/2}|\widehat{QLR}_n|/\widehat{\omega}_n > z_{\alpha/2}\big) = AsySZ^{no}(\alpha). \qquad \text{(B.11)}$$

By part (a), the l.h.s. of the equation above equals α. Therefore, $AsySZ^{no}(\alpha) = \alpha$. □

Proof of Theorem 2. In this proof, we focus on the two-sided test. Arguments for the one-sided test is the same.

(a) Let $\{a_n\}$ be a subsequence of $\{n\}$ such that $AsySZ^{ol}(\alpha) = \lim_{n\to\infty} SZ_{a_n}^{ol}(\alpha)$. Such a sequence always exists. Let $\{\mu_n \in \mathcal{H}_0^{ol}\}_{n=1}^{\infty}$ be a sequence such that for each n,

$$\Pr_n\big(n^{1/2}|\widehat{QLR}_n|/(\widehat{\omega}_n \vee n^{-1/2}b_n) > z_{\alpha/2}\big) \geq SZ_n^{ol}(\alpha) - o(1). \qquad \text{(B.12)}$$

Let $\{u_n\}$ be a subsequence of $\{a_n\}$ such that $u_n\omega_{u_n}^2 \to \sigma$, $\sigma \in [0, \infty]$. Such subsequences always exist because we allow σ to take values in the extended real space. Then,

$$AsySZ^{ol}(\alpha) = \lim_{n\to\infty}\Pr_{u_n}\big(u_n^{1/2}|\widehat{QLR}_{u_n}|/(\widehat{\omega}_{u_n} \vee u_n^{-1/2}b_{u_n}) > z_{\alpha/2}\big). \qquad \text{(B.13)}$$

If $\sigma < \infty$, then by Lemma 6(a) and $b_n \to \infty$,

$$\begin{aligned}
&\lim_{n\to\infty}\Pr_{u_n}\big(u_n^{1/2}|\widehat{QLR}_{u_n}|/(\widehat{\omega}_{u_n} \vee u_n^{-1/2}b_{u_n}) > z_{\alpha/2}\big)\\
&\quad \leq \lim_{n\to\infty}\Pr_{u_n}\big(u_n|\widehat{QLR}_{u_n}| > b_{u_n}z_{\alpha/2}\big) = 0 < \alpha. \qquad \text{(B.14)}
\end{aligned}$$

If $\sigma = \infty$, then by Lemma 6(b) and $n^{-1/2}b_n \to 0$,

$$\begin{aligned}
&\lim_{n\to\infty}\Pr_{u_n}\big(u_n^{1/2}|\widehat{QLR}_{u_n}|/(\widehat{\omega}_{u_n} \vee u_n^{-1/2}b_{u_n}) > z_{\alpha/2}\big)\\
&\quad \leq \lim_{n\to\infty}\Pr_{u_n}\big(u_n^{1/2}|\widehat{QLR}_{u_n}|/\widehat{\omega}_{u_n} > z_{\alpha/2}\big) = \alpha. \qquad \text{(B.15)}
\end{aligned}$$

Therefore, by (B.13)–(B.15), $AsySZ^{ol}(\alpha) \leq \alpha$.

(b) Let $\mu \in \mathcal{H}_0^{ol}$ satisfy $P_\mu^* \neq Q_\mu^*$. By Assumption 6, such a μ exists. Then, $\omega_\mu^2 > 0$. By Lemma 6(b), under μ, $\widehat{\omega}_n^2 \to_p \omega_\mu^2 > 0$. Also, by Lemma 6(b), under μ, $n^{1/2}\widehat{QLR}_n/\widehat{\omega}_n \to_d N(0,1)$. Because $n^{-1/2}b_n \to 0$, we have

$$\lim_{n\to\infty} \Pr_\mu\left(n^{1/2}|\widehat{QLR}_n|/(\widehat{\omega}_n \vee n^{-1/2}b_n) > z_{\alpha/2}\right) = \alpha. \tag{B.16}$$

By definition, $AsySZ^{ol}(\alpha) \geq \lim_{n\to\infty}\Pr_\mu\left(n^{1/2}|\widehat{QLR}_n|/(\widehat{\omega}_n \vee n^{-1/2}b_n) > z_{\alpha/2}\right)$. Thus, we have $AsySZ^{ol}(\alpha) \geq \alpha$. Combining this with part (a), we obtain the desired result. □

Proof of Theorem 3. First by Lemma A.3(b) and condition (iii) Definition H, we have $\{\hat{\gamma}_n(\theta) : \theta \in \Theta\} \subseteq \Gamma_M^m$ and $\{\hat{\lambda}_n(\beta) : \beta \in B\} \subseteq \Gamma_M^g$ with probability approaching one. Thus, with probability approaching one,

$$\begin{aligned}\sqrt{n}\widehat{QLR}_n &= \sqrt{n}(\max_{\theta\in\Theta}\min_{\gamma\in\Gamma_M^m}\widehat{\mathcal{M}}_n(\gamma,\theta) \\ &\quad - \max_{\beta\in B}\min_{\lambda\in\Gamma_M^g}\widehat{\mathcal{N}}_n(\lambda,\beta)).\end{aligned} \tag{B.17}$$

Rewrite the r.h.s. by adding and subtracting terms, and we get

$$\begin{aligned}\sqrt{n}\widehat{QLR}_n &= \sqrt{n}(\max_{\theta\in\Theta}\min_{\gamma\in\Gamma_M^m}\widehat{\mathcal{M}}_n(\gamma,\theta) - \max_{\theta\in\Theta}\min_{\gamma\in\Gamma_M^m}\mathcal{M}_\mu(\gamma,\theta)) \\ &\quad - \sqrt{n}(\max_{\beta\in B}\min_{\lambda\in\Gamma_M^g}\widehat{\mathcal{N}}_n(\lambda,\beta) - \max_{\beta\in B}\min_{\lambda\in\Gamma_M^g}\mathcal{N}_\mu(\lambda,\beta)) \\ &\quad + \sqrt{n}(\mathcal{M}_\mu^* - \mathcal{N}_\mu^*),\end{aligned} \tag{B.18}$$

where the first equality holds by Lemma A.3(b) and condition (iii) Definition H. By Lemma A.2(a), we have

$$\begin{aligned}&\left|\sqrt{n}(\max_{\theta\in\Theta}\min_{\gamma\in\Gamma_M^m}\widehat{\mathcal{M}}_n(\gamma,\theta) - \max_{\theta\in\Theta}\min_{\gamma\in\Gamma_M^m}\mathcal{M}_\mu(\gamma,\theta))\right| \\ &\quad\leq \sup_{\phi\in\Theta\times\Gamma_M^m}\left|\nu_n^0(\phi)\right| = O_p(1).\end{aligned} \tag{B.19}$$

Similarly, $\left|\sqrt{n}(\max_{\beta\in B}\min_{\lambda\in\Gamma_M^g}\widehat{\mathcal{N}}_n(\lambda,\beta) - \max_{\beta\in B}\min_{\lambda\in\Gamma_M^g}\mathcal{N}_\mu(\lambda,\beta))\right| = O_p(1)$. Also, by Lemma 2(b),

$$\begin{aligned}\sqrt{n}(\mathcal{M}_\mu^* - \mathcal{N}_\mu^*) &= \sqrt{n}(\exp(-d(\mathcal{P},\mu)) - \exp(-d(\mathcal{Q},\mu))) \\ &\to \infty.\end{aligned} \tag{B.20}$$

Therefore, for any $C > 0$, $\lim_{n\to\infty}\Pr_\mu(\sqrt{n}\widehat{QLR}_n > C) = 1$.

Now for the denominator $\widehat{\omega}_n$, we have

$$\begin{aligned}E_\mu[\widehat{\omega}_n^2] &\leq E_\mu\Bigg[\sup_{(\theta',\gamma',\beta',\lambda')\in\Theta\times\Gamma_M^m\times B\times\Gamma_M^g} n^{-1}\sum_{i=1}^n\big(\exp(\gamma' m_i(\theta)) \\ &\qquad - \exp(\lambda' g_i(\beta))\big)^2\Bigg] \\ &\leq E_\mu\Bigg[2\sup_{(\theta',\gamma',\beta',\lambda')'\in\Theta\times\Gamma_M^m\times B\times\Gamma_M^g} n^{-1}\sum_{i=1}^n\big(\exp(2\gamma' m_i(\theta)) \\ &\qquad + \exp(2\lambda' g_i(\beta))\big)\Bigg] \\ &\leq 2n^{-1}\sum_{i=1}^n\Bigg(E_\mu\sup_{(\theta',\gamma')'\in\Theta\times\Gamma_M^m}\exp(2\gamma' m_i(\theta)) \\ &\qquad + E_\mu\sup_{(\beta',\lambda')'\in B\times\Gamma_M^g}\exp(2\lambda' g_i(\beta))\Bigg) \\ &\leq 2M,\end{aligned} \tag{B.21}$$

where last inequality holds by condition (v) of Definition H. Therefore, $\widehat{\omega}_n^2 = O_p(1)$.

Therefore, for any $C > 0$, $\lim_{n\to\infty}\Pr_\mu(\sqrt{n}\widehat{QLR}_n/\widehat{\omega}_n > C) = 1$. This shows part (a).

Part (b) follows because $\widehat{\omega}_n \vee (n^{-1/2}b_n) \geq \widehat{\omega}_n$. □

Proof of Theorem 4. The proof is the same as that for Theorem 3 except with μ_n in place of μ and with (B.20) modified as follows:

$$\begin{aligned}\sqrt{n}(\mathcal{M}_{\mu_n}^* - \mathcal{N}_\mu^*) &= \exp(-\tilde{d}_n)\sqrt{n}(-d(\mathcal{P},\mu_n) + d(\mathcal{Q},\mu_n)) \\ &= -(\exp(-d(\mathcal{P},\mu_0)) + o_p(1))\sqrt{n}(d(\mathcal{P},\mu_n) - d(\mathcal{Q},\mu_n)),\end{aligned} \tag{B.22}$$

where $\tilde{d}_n$ lies in between $d(\mathcal{P},\mu_n)$ and $d(\mathcal{Q},\mu_n)$. Because $d(\mathcal{P},\mu_0) < \infty$, $\exp(-d(\mathcal{P},\mu_0)) + o(1) > \varepsilon$ eventually as $n \to \infty$ for some $\varepsilon > 0$. Therefore,

$$\sqrt{n}(\mathcal{M}_{\mu_n}^* - \mathcal{N}_{\mu_n}^*) \to \infty. \quad \square \tag{B.23}$$

Appendix C. Proof of the main lemmas

Proof of Lemma 1. We only need to show parts (a)–(b) because part (c) is analogous.

(a) By Assumption 1(a)–(b), $p_{\theta,\mu}^*$ is a well defined density function. The proof here is similar to that of the second part of Theorem 3.1 in Csiszár (1975). Let P be a distribution in $\mathcal{P}_\theta$ such that $P \ll \mu$ and let p_μ denote the density of P with respect to μ, then

$$\begin{aligned}d(P,\mu) - d(P,P_{\theta,\mu}^*) &= \int\log p_\mu dP - \int\log(p_\mu/p_{\theta,\mu}^*)dP \\ &= \int\log p_{\theta,\mu}^* dP \\ &= -\log\left[E_\mu\exp(\gamma_\mu^*(\theta)'m(X_i,\theta))\right] \\ &\quad + \gamma_\mu^*(\theta)'E_P m(X_i,\theta) \\ &\geq -\log E_\mu\exp(\gamma_\mu^*(\theta)'m(X_i,\theta)),\end{aligned} \tag{C.1}$$

where the inequality holds because for $j \leq d_p$, $E_P m_j(X_i,\theta) = 0$, and for $j \geq d_p + 1$, $E_P m_j(X_i,\theta) \geq 0$ and $\gamma_{\mu,j}^*(\theta) \geq 0$. Eq. (C.1) implies that

$$d(P,\mu) \geq -\log E_\mu\exp(\gamma_\mu^*(\theta)'m(X_i,\theta)). \tag{C.2}$$

By definition,

$$\begin{aligned}d(P_{\theta,\mu}^*,\mu) &= \int\log p_{\theta,\mu}^* dP_\theta^* \\ &= -\log E_\mu\exp(\gamma_\mu^*(\theta)'m(X_i,\theta)) + \gamma_\mu^*(\theta)'E_{P_\theta^*}m(X_i,\theta) \\ &= -\log E_\mu\exp(\gamma_\mu^*(\theta)'m(X_i,\theta)),\end{aligned} \tag{C.3}$$

where the last equality holds by the Kuhn–Tucker conditions from the minimization problem: $\min_{\gamma\in R^{dp}\times R_+^{dm-dp}}\mathcal{M}_\mu(\gamma,\theta)$. The Kuhn–Tucker conditions are

$$\begin{aligned}&0 = \partial\mathcal{M}_\mu(\gamma_\mu^*(\theta),\theta)/\partial\gamma_j \equiv E_{P_\theta^*}m_j(X_i,\theta) \quad \text{for } j \leq d_p \\ &0 = \gamma_{\mu,j}^*(\theta)\big(\partial\mathcal{M}_\mu(\gamma_\mu^*(\theta),\theta)/\partial\gamma_j\big) \equiv \gamma_{\mu,j}^*(\theta)E_{P_\theta^*}m_j(X_i,\theta) \\ &\quad \text{for } j \geq d_p + 1.\end{aligned} \tag{C.4}$$

By (C.2) and (C.3), we have $d(P^*_{\theta,\mu},\mu) = \min_{P\in\mathcal{P}_\theta} d(P,\mu)$, that is, $P^*_{\theta,\mu}$ is the I-projection of μ on P.

(b) Part (b) is implied by (C.3). □

Proof of Lemma 2. (a) By Assumptions 1(a)–(b) and 2(b), $\gamma^*_\mu(\theta)$ is the unique minimizer of the function $\mathcal{M}_\mu(\gamma,\theta)$ and $\mathcal{M}_\mu(\gamma,\theta)$ is continuous in (γ,θ). The maximum theorem then implies that $\gamma^*_\mu(\theta)$ is continuous in θ. Consequently, $\mathcal{M}_\mu(\gamma^*_\mu(\theta),\theta)$ is continuous in θ. The continuity of $\mathcal{M}_\mu(\gamma^*_\mu(\theta),\theta)$ combined with Assumption 2(a) implies part (a)

(b) By part (a), $\sup_{\theta\in\Theta}\mathcal{M}_\mu(\gamma^*_\mu(\theta),\theta) = \max_{\theta\in\Theta}\mathcal{M}_\mu(\gamma^*_\mu(\theta),\theta)$. By Lemma 1(b) and the definition of $\gamma^*_\mu(\cdot)$, we have part (b).

(c) The arguments for part (c) are analogous to those for parts (a)–(b). □

Proof of Lemma 3. It suffices to show $\rho_{lh}(\widehat{\Theta}_n,\Theta^*_n)\to_p 0$ because $\rho_{lh}(\widehat{B}_n,B^*_n)\to_p 0$ can be obtained by analogous arguments.

For an arbitrary $\varepsilon>0$ and an arbitrary sequence $\{\hat\theta_n\in\widehat{\Theta}_n\}_{n=1}^\infty$, and arbitrary $\theta^*_n\in\Theta^*_n$,

$$\begin{aligned}
&\Pr_n\big(\rho_{lh}(\widehat{\Theta}_n,\Theta^*_n)>\varepsilon\big)\\
&\quad\le \Pr_n\big(\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\hat\theta_n))>\delta_\varepsilon\big)\\
&\quad= \Pr_n\big([\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))-\widehat{\mathcal{M}}_n(\hat\phi_n(\theta^*_n))]\\
&\qquad+[\widehat{\mathcal{M}}_n(\hat\phi_n(\theta^*_n))-\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))]\\
&\qquad+[\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\hat\theta_n))]>\delta_\varepsilon\big)\\
&\quad= \Pr_n\big(o_p(1)+\widehat{\mathcal{M}}_n(\hat\phi_n(\theta^*_n))-\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))+o_p(1)>\delta_\varepsilon\big)\\
&\quad\le \Pr_n(o_p(1)+o_p(1)>\delta_\varepsilon)\to 0,
\end{aligned}\tag{C.5}$$

where the first inequality holds by condition (ii) in (5.4), the second equality holds by Lemmas A.2(e) and A.3(a), and the second inequality holds because $\hat\theta_n$ maximizes $\widehat{\mathcal{M}}_n(\hat\gamma_n(\theta),\theta)$. Thus, $\rho_{lh}(\widehat{\Theta}_n,\Theta^*_n)\to_p 0$. □

Proof of Lemma 4. The lemma follows from the derivation belows. For any $\mu\in\mathcal{H}_0^{no}$ and $\theta^*\in\Theta^*_\mu$, we have

$$\begin{aligned}
\omega^2_\mu &= \mathcal{M}^2_\mu(\gamma^*_\mu(\theta^*),\theta^*)E_\mu[dP^*_\mu/d\mu-dQ^*_\mu/d\mu]^2\\
&\ge E_\mu[dP^*_\mu/d\mu-dQ^*_\mu/d\mu]^2\cdot\exp(-2M_1)\\
&\ge \big[E_\mu|dP^*_\mu/d\mu-dQ^*_\mu/d\mu|\big]^2\cdot\exp(-2M_1)\\
&= \left[\int|dP^*_\mu/d\nu_{P^*_\mu,Q^*_\mu}-dQ^*_\mu/d\nu_{P^*_\mu,Q^*_\mu}|d\nu_{P^*_\mu,Q^*_\mu}\right]^2\cdot\exp(-2M_1)\\
&\ge \inf_{P\in\mathcal{P},Q\in\mathcal{Q}}\left[\int|dP/d\nu_{P,Q}-dQ/d\nu_{P,Q}|d\nu_{P,Q}\right]^2\\
&\quad\cdot\exp(-2M_1)>0,
\end{aligned}\tag{C.6}$$

where the first equality holds by Lemma 1(a) and Definition H0NO ($\mathcal{M}_\mu(\gamma^*_\mu(\theta^*),\theta^*)=\mathcal{N}_\mu(\lambda^*_\mu(\beta^*),\beta^*)$), the first inequality holds because $\mathcal{M}_\mu(\gamma^*_\mu(\theta^*),\theta^*)=\exp(-d(\mathcal{P},\mu))\ge\exp(-M_1)$ by Lemma 2(b) and condition (i) in Definition H0NO, the second inequality holds by the convexity of $f(x)=x^2$, the second equality holds because P^*_μ and Q^*_μ are absolutely continuous w.r.t. $\nu_{P^*_\mu,Q^*_\mu}$, the third inequality holds because $P^*_\mu\in\mathcal{P}$ and $Q^*_\mu\in\mathcal{Q}$, and the last inequality holds by Definition NO. □

Proof of Lemma 5. It suffices to show that $\rho_{lh}(\widehat{\Theta}_n,\Theta^*_n)=O_p(n^{-1/2})$ because the remainder is analogous. We use the consistency already shown in Lemma 3: $\rho_{lh}(\widehat{\Theta}_n,\Theta^*_n)\to_p 0$.

Take an arbitrary sequence $\{\hat\theta_n\in\widehat{\Theta}_n\}_{n=1}^\infty$. Let $\{\theta^*_n\in\Theta^*_n\}_{n=1}^\infty$ be a sequence such that $\|\theta^*_n-\hat\theta_n\|^2\le\rho^2_{lh}(\hat\theta_n,\Theta^*_n)+o(n^{-1/2})$. The proof is based on the quadratic approximation of $\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))-\widehat{\mathcal{M}}_n(\phi^*_n(\theta^*_n))$ and that of $\mathcal{M}_{\mu_n}(\hat\phi_n(\hat\theta_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))$. The basic idea is from Andrews (1999), but the procedure is more involved here because (a) we deal with a saddle-point estimation problem instead of a extremum estimation problem, (b) after profiling out the first step minimization parameter γ, the criterion functions $\widehat{\mathcal{M}}_n(\hat\phi_n(\theta))$ and $\mathcal{M}_{\mu_n}(\phi^*_n(\theta))$ are non-differentiable in θ, and (c) there is no straightforward way of writing down the left/right derivatives w.r.t. θ. We construct quadratic bounds for the centralized population and sample criterion functions. Specifically, we show below that

$$\begin{aligned}
&\text{(i) }\big[\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))-\widehat{\mathcal{M}}_n(\phi^*_n(\theta^*_n))\big]\\
&\qquad-\big[\mathcal{M}_{\mu_n}(\hat\phi_n(\hat\theta_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))\big]\\
&\quad=O_p(n^{-1})+O_p(n^{-1/2})\cdot\|\hat\theta_n-\theta^*_n\|\\
&\qquad+o_p(1)\cdot\|\hat\theta_n-\theta^*_n\|^2,\\
&\text{(ii) }\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))-\widehat{\mathcal{M}}_n(\phi^*_n(\theta^*_n))\ge O_p(n^{-1}),\quad\text{and}\\
&\text{(iii) }\mathcal{M}_{\mu_n}(\hat\phi_n(\hat\theta_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))\le O_p(n^{-1})-C\\
&\qquad\cdot\big((\|\hat\theta_n-\theta^*_n\|^2-o(n^{-1}))\wedge\delta\big),
\end{aligned}\tag{C.7}$$

where C and δ are the positive constants in condition (iv) of Definition H0OL. Conditions (i)–(iii) in (C.7) imply that

$$\begin{aligned}
O_p(n^{-1})&\le O_p(n^{-1/2})\cdot\|\hat\theta_n-\theta^*_n\|+o_p(1)\cdot\|\hat\theta_n-\theta^*_n\|^2\\
&\quad-C\cdot\big((\|\hat\theta_n-\theta^*_n\|^2-o(n^{-1}))\wedge\delta\big)\\
&=-C\|\hat\theta_n-\theta^*_n\|^2+O_p(n^{-1/2})\cdot\|\hat\theta_n-\theta^*_n\|\\
&\quad+o_p(1)\cdot\|\hat\theta_n-\theta^*_n\|^2+C\cdot o(n^{-1}),
\end{aligned}\tag{C.8}$$

where the equality holds with probability approaching one because $\|\hat\theta_n-\theta^*_n\|^2-o(n^{-1})\to_p 0$ by Lemma 3. The above equation implies that $\|\hat\theta_n-\theta^*_n\|=O_p(n^{-1/2})$. Therefore, the desired result, $\rho_{lh}(\hat\Theta_n,\Theta^*_n)=O_p(n^{-1/2})$, holds since $\hat\theta_n$ is arbitrarily chosen from $\widehat{\Theta}_n$.

Now, we show condition (i) in (C.7). We have

$$\begin{aligned}
&\big[\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))-\widehat{\mathcal{M}}_n(\phi^*_n(\theta^*_n))\big]-\big[\mathcal{M}_{\mu_n}(\hat\phi_n(\hat\theta_n))-\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))\big]\\
&\quad=\big[\partial\widehat{\mathcal{M}}_n(\phi^*_n(\theta^*_n))/\partial\phi'-\partial\mathcal{M}_{\mu_n}(\phi^*_n(\theta^*_n))/\partial\phi'\big]\\
&\qquad\times\big[\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\big]\\
&\qquad+2^{-1}\big[\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\big]'\big[\partial^2\widehat{\mathcal{M}}_n(\tilde\phi_n)/\partial\phi\partial\phi'\\
&\qquad-\partial^2\mathcal{M}_{\mu_n}(\bar\phi_n)/\partial\phi\partial\phi'\big]\big[\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\big]\\
&\quad=O_p(n^{-1/2})\cdot\|\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\|\\
&\qquad+o_p(1)\cdot\|\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\|^2
\end{aligned}\tag{C.9}$$

where both $\tilde\phi_n$ and $\bar\phi_n$ lie on the line segment joining $\hat\phi_n(\hat\theta_n)$ and $\phi^*_n(\theta^*_n)$ and they are not necessarily the same, the first equality holds by second-order Taylor expansions of $\widehat{\mathcal{M}}_n(\hat\phi_n(\hat\theta_n))$ and $\mathcal{M}_{\mu_n}(\hat\phi_n(\hat\theta_n))$ around $\phi^*_n(\beta^*_n)$, the second equality holds by Lemmas A.2(d) and A.4(a). Now observe that

$$\begin{aligned}
\|\hat\phi_n(\hat\theta_n)-\phi^*_n(\theta^*_n)\|^2&=\|\hat\gamma_n(\hat\theta_n)-\gamma^*_n(\theta^*_n)\|^2+\|\hat\theta_n-\theta^*_n\|^2\\
&\le 2\|\hat\gamma_n(\hat\theta_n)-\gamma^*_n(\hat\theta_n)\|^2\\
&\quad+2\|\gamma^*_n(\hat\theta_n)-\gamma^*_n(\theta^*_n)\|^2+\|\hat\theta_n-\theta^*_n\|^2\\
&\le(\sqrt{2}\|\hat\gamma_n(\hat\theta_n)-\gamma^*_n(\hat\theta_n)\|\\
&\quad+\sqrt{2}\|\gamma^*_n(\hat\theta_n)-\gamma^*_n(\theta^*_n)\|+\|\hat\theta_n-\theta^*_n\|)^2\\
&=(O_p(n^{-1/2})+O_p(\|\hat\theta_n-\theta^*_n\|))^2
\end{aligned}\tag{C.10}$$

where the first inequality holds by the triangular inequality and the convexity of the square function, the second inequality holds by $(a+b+c)^2\ge a^2+b^2+c^2$ for $a,b,c\ge 0$ and the equality holds by Lemma A.3(b)–(c). This combined with Eq. (C.9) implies condition (i) in (C.7).

Condition (ii) in (C.7) is implied by

$$\begin{aligned}
&\widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*)) \geq \widehat{\mathcal{M}}_n(\hat{\phi}_n(\theta_n^*)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*)) \\
&\quad = [\partial \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))/\partial\gamma'][\hat{\gamma}_n(\theta_n^*) - \gamma_n^*(\theta_n^*)] \\
&\qquad + 2^{-1}\left[\hat{\gamma}_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\right]'[\partial^2 \widehat{\mathcal{M}}_n(\tilde{\phi}_n)\partial\gamma\partial\gamma'] \\
&\qquad \times \left[\hat{\gamma}_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\right] \\
&\quad = [\partial \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))/\partial\gamma'][\hat{\gamma}_n(\theta_n^*) - \gamma_n^*(\theta_n^*)] + O_p(n^{-1}) \\
&\quad \geq \left[\partial \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))/\partial\gamma' - \partial \mathcal{M}_{\mu_n}(\phi_n^*(\theta_n^*))/\partial\gamma'\right] \\
&\qquad \times \left[\hat{\gamma}_n(\theta_n^*) - \gamma_n^*(\theta_n^*)\right] + O_p(n^{-1}) \\
&\quad = O_p(n^{-1}),
\end{aligned} \tag{C.11}$$

where $\tilde{\phi}_n$ lies on the line segment joining $\hat{\phi}_n(\theta_n^*)$ and $\phi_n^*(\theta_n^*)$, the first inequality holds because $\widehat{\Theta}_n$ is a maximizer of $\widehat{\mathcal{M}}_n(\hat{\phi}_n(\theta))$, the first equality holds by a Taylor expansion, the second equality holds by Lemmas A.2(e) and A.3(b) and condition (v) of Definition H, the second inequality holds by the same arguments as those for the second inequality in (B.6) and the last equality holds by Lemmas A.2(d) and A.3(b).

Condition (iii) in (C.7) is implied by

$$\begin{aligned}
&\mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n)) - \mathcal{M}_{\mu_n}(\phi_n^*(\theta_n^*)) = \left[\mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n)) - \mathcal{M}_{\mu_n}(\phi_n^*(\hat{\theta}_n))\right] \\
&\qquad + \left[\mathcal{M}_{\mu_n}(\phi_n^*(\hat{\theta}_n)) - \mathcal{M}_{\mu_n}(\phi_n^*(\theta_n^*))\right] \\
&\quad \leq \left[\mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n)) - \mathcal{M}_{\mu_n}(\phi_n^*(\hat{\theta}_n))\right] \\
&\qquad - C \cdot \left((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1})) \wedge \delta\right) \\
&\quad = -C \cdot \left((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1})) \wedge \delta\right) \\
&\qquad + [\partial \mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma'][\hat{\gamma}_n(\hat{\theta}_n) - \gamma_n^*(\hat{\theta}_n)] \\
&\qquad - 2^{-1}\left[\hat{\gamma}_n(\hat{\theta}_n) - \gamma_n^*(\hat{\theta}_n)\right]'[\partial^2 \mathcal{M}_{\mu_n}(\tilde{\phi}_n)\partial\gamma\partial\gamma'] \\
&\qquad \times \left[\hat{\gamma}_n(\hat{\theta}_n) - \gamma_n^*(\hat{\theta}_n)\right] \\
&\quad = O_p(n^{-1}) - C \cdot \left((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1})) \wedge \delta\right) \\
&\qquad + [\partial \mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma']\left[\hat{\gamma}_n(\hat{\theta}_n) - \gamma_n^*(\hat{\theta}_n)\right] \\
&\quad \leq O_p(n^{-1}) - C \cdot \left((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1})) \wedge \delta\right) \\
&\qquad + \left[\partial \mathcal{M}_{\mu_n}(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma' - \partial \widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma'\right] \\
&\qquad \times \left[\hat{\gamma}_n(\hat{\theta}_n) - \gamma_n^*(\hat{\theta}_n)\right] \\
&\quad = O_p(n^{-1}) - C \cdot \left((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1})) \wedge \delta\right),
\end{aligned} \tag{C.12}$$

where $\tilde{\phi}_n$ lies on the line segment joining $\phi_n^*(\hat{\theta}_n)$ and $\hat{\phi}_n(\hat{\theta}_n)$, the first inequality holds by condition (iv) of Definition H0OL and $\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1}) \leq \rho_{lh}^2(\widehat{\Theta}_n, \Theta_n^*)$ by design, the second equality holds by a Taylor expansion of $\mathcal{M}_{\mu_n}(\phi_n^*(\hat{\theta}_n))$ around $\hat{\phi}_n(\hat{\theta}_n)$, the third equality holds by Lemmas A.2(e) and A.3(b), the second inequality holds by $(\partial \widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma')\hat{\gamma}_n(\hat{\theta}_n) = 0$ and $\partial \widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n))/\partial\gamma_j \begin{cases} = 0 & \text{for } j \leq d_p \\ \geq 0 & \text{for } j > d_p \end{cases}$ both being the Kuhn–Tucker conditions of the minimization problem: $\min_{\gamma \in R^{d_p} \times R_+^{d_m - d_p}} \widehat{\mathcal{M}}_n(\gamma, \hat{\theta}_n)$, and the last equality holds by Lemmas A.2(d) and A.3(b). □

Proof of Lemma 6. The lemma is stated in terms of subsequences $\{u_n\}_{n=1}^\infty$. For notational simplicity, we prove it for the sequence $\{n\}$. All of the arguments go through with $\{u_n\}$ in place of $\{n\}$. Let $\hat{\theta}_n \in \widehat{\Theta}_n$ and $\hat{\beta}_n \in \widehat{B}_n$ be those that satisfy $\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n) = \widehat{\omega}_n^2$.

(a) Let $\{\theta_n^* \in \Theta_n^*\}_{n=1}^\infty$ be a sequence such that $\|\theta_n^* - \hat{\theta}_n\|^2 \leq \rho_{lh}^2(\widehat{\Theta}_n, \Theta_n^*) + o(n^{-1})$. Then by Lemma 5, $\|\theta_n^* - \hat{\theta}_n\| = O_p(n^{-1/2})$. We first show $n\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n) = O_p(1)$. Observe that

$$\begin{aligned}
&n\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n) \leq \sum_{i=1}^n (e^{\hat{\gamma}_n m_i(\hat{\theta}_n)} - e^{\hat{\lambda}_n g_i(\hat{\beta}_n)})^2 \\
&\quad \leq 3\sum_{i=1}^n (\Lambda_{n,i}^*)^2 + 3n(\hat{\phi}_n(\hat{\theta}_n) - \phi_n^*(\theta_n^*))' \\
&\quad \times \left(n^{-1}\sum_{i=1}^n \frac{\partial e^{\tilde{\gamma}_n' m_i(\tilde{\theta}_n)}}{\partial\phi}\frac{\partial e^{\tilde{\gamma}_n' m_i(\tilde{\theta}_n)}}{\partial\phi'}\right)(\hat{\phi}_n(\hat{\theta}_n) - \phi_n^*(\theta_n^*)) \\
&\quad + 3n(\hat{\psi}_n(\hat{\beta}_n) - \psi_n^*(\beta_n^*))'\left(n^{-1}\sum_{i=1}^n \frac{\partial e^{\tilde{\lambda}_n' g_i(\tilde{\beta}_n)}}{\partial\psi}\frac{\partial e^{\tilde{\lambda}_n' g_i(\tilde{\beta}_n)}}{\partial\psi'}\right) \\
&\quad \times (\hat{\psi}_n(\hat{\beta}_n) - \psi_n^*(\beta_n^*)) \\
&\quad \equiv 3(W_{n,1} + W_{n,2} + W_{n,3}),
\end{aligned} \tag{C.13}$$

where $\tilde{\phi}_n$ and $\tilde{\psi}_n$ lie on the lie segment joining $\hat{\phi}_n(\hat{\theta}_n)$ and $\phi_n^*(\theta_n^*)$ and the one joining $\hat{\psi}_n(\hat{\beta}_n)$ and $\psi_n^*(\beta_n^*)$, respectively, the second inequality holds by a mean-value expansion and the inequality $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$. In (C.13), $W_{n,1} = O_p(1)$ because $E_n\left|W_{n,1}\right| = n\omega_n^2 \to \sigma^2 < \infty$. Also, $W_{n,2} = O_p(1)$ by (B.10), (C.10) and Lemma 5. Finally, $W_{n,3} = O_p(1)$ for analogous reasons. Therefore, $n\hat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n) = O_p(1)$ when $\sigma < \infty$.

Now we show $n\widehat{QLR}_n = O_p(1)$. Observe that

$$\begin{aligned}
n\widehat{QLR}_n &= \sum_{i=1}^n \Lambda_{n,i}^* + n\left(\widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))\right) \\
&\quad - n\left(\widehat{\mathcal{N}}_n(\hat{\psi}_n(\hat{\beta}_n)) - \widehat{\mathcal{N}}_n(\psi_n^*(\beta_n^*))\right) \\
&= O_p(1) + n\left(\widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))\right) \\
&\quad - n\left(\widehat{\mathcal{N}}_n(\hat{\psi}_n(\hat{\beta}_n)) - \widehat{\mathcal{N}}_n(\psi_n^*(\beta_n^*))\right) \\
&= O_p(1) + O_p(1) - n\left(\widehat{\mathcal{N}}_n(\hat{\psi}_n(\hat{\beta}_n)) - \widehat{\mathcal{N}}_n(\psi_n^*\,(\beta_n^*))\right) \\
&= O_p(1) + O_p(1) - O_p(1) = O_p(1),
\end{aligned} \tag{C.14}$$

where the second equality holds because $E_n\left(\sum_{i=1}^n \Lambda_{n,i}^*\right)^2 = \sum_{i=1}^n \left(E_n\Lambda_{n,i}^*\right)^2 = n\omega_n^2 \to \sigma^2 < \infty$, the third equality holds by (C.15), and the fourth equality holds for analogous reasons as the third. Therefore, $n\widehat{QLR}_n = O_p(1)$.

(b) The proof here is of the same structure as, but slightly different from, the proof of Theorem 1(a). The difference is caused by the fact that (i) ω_n^2 is not bounded away from zero in this lemma while it is under the conditions of Theorem 1(a), and (ii) the set estimators are $n^{-1/2}$-consistent in this lemma while they are not in Theorem 1(a).

First, we show $n^{1/2}\widehat{QLR}_n/\omega_n \to_d N(0, 1)$. Let $A_{n,1}$ and $A_{n,2}$ be the same as in (B.4). Then, by (B.4), the desired result is implied by (i) $\omega_n^{-1}n^{-1/2}\sum_{i=1}^n \Lambda_{n,i}^* \to_d N(0, 1)$, (ii) $A_{n,1} = o_p(1)$ and (iii) $A_{n,2} = o_p(1)$. Conditions (i)–(ii) are shown below. Condition (iii) holds for analogous reasons as condition (ii).

By the Lyapunov CLT, (i) holds. The CLT applies because (a) $E_n\Lambda_{n,i}^* = 0$ by condition (ii) of Definition H0OL and Lemma 2(b), (b) $\omega_n^{-2}E_n(\Lambda_{n,i}^*)^2 = 1$, and (c) $E_n(\omega_n^{-1}\Lambda_{n,i}^*)^{2+\delta} < \infty$ by condition (iii) of Definition H0OL.

Now we show (ii) $A_{n,1} = o_p(1)$. Because $A_{n,1} = n^{1/2}\omega_n^{-1}\left(\widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*))\right)$, (ii) is implied by $n\omega_n^2 \to \infty$ and the following derivation:

$$\begin{aligned}
O_p(n^{-1}) &\leq \widehat{\mathcal{M}}_n(\hat{\phi}_n(\hat{\theta}_n)) - \widehat{\mathcal{M}}_n(\phi_n^*(\theta_n^*)) \\
&\leq O_p(n^{-1/2}) \cdot \|\hat{\theta}_n - \theta_n^*\| + o_p(1) \cdot \|\hat{\theta}_n - \theta_n^*\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ O_p(n^{-1}) - C \cdot ((\|\hat{\theta}_n - \theta_n^*\|^2 - o(n^{-1/2})) \wedge \delta) \\
&= O_p(n^{-1}), \qquad \text{(C.15)}
\end{aligned}
$$

where the first inequality holds by condition (ii) in (C.7) in the proof of Lemma 5, the second inequality holds by conditions (i) and (iii) in (C.7), and the equality holds by Lemma 5.

Next, we show $\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n)/\omega_n^2 \to_p 1$. Decompose $\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n)/\omega_n^2$ in the same way as in (B.8). We show below that $W_{n,0} = o_p(1)$, $W_{n,1} \to_p 1$, $W_{n,2} = o_p(1)$ and $W_{n,3} = o_p(1)$ when $\sigma = \infty$. These results together imply $\widehat{\omega}_n^2(\hat{\theta}_n, \hat{\beta}_n)/\omega_n^2 \to_p 1$.

The first summand $W_{n,0} \to_p 0$ by $n^{1/2}\widehat{QLR}_n/\omega_n \to_d N(0, 1)$. The second summand $W_{n,1} \equiv \omega_n^{-2} n^{-1} \sum_{i=1}^n [\Lambda_{n,i}^*]^2 \to_p 1$ by LLN. The LLN applies because (a) $E_n \omega_n^{-2} [\Lambda_{n,i}^*]^2 = 1$ and (b) $\sup_{n \geq 1} E_n [\omega_n^{-1} \Lambda_{n,i}^*]^{2+\delta} < \infty$ by condition (ii) of Definition H0OL.

The term $W_{n,3}$ is $o_p(1)$ because

$$
\begin{aligned}
0 &\leq W_{n,3} \\
&\leq 2\omega_n^{-2} (\hat{\phi}_n(\hat{\theta}_n) - \phi_n^*(\theta_n^*)) \left(n^{-1} \sum_{i=1}^n \frac{\partial e^{\tilde{\gamma}_n' m_i(\tilde{\theta}_n)}}{\partial \phi} \frac{\partial e^{\tilde{\gamma}_n' m_i(\tilde{\theta}_n)}}{\partial \phi'} \right) \\
&\quad \times (\hat{\phi}_n(\hat{\theta}_n) - \phi_n^*(\theta_n^*)) \\
&\quad + 2\omega_n^{-2} (\hat{\psi}_n(\hat{\beta}_n) - \psi_n^*(\beta_n^*)) \left(n^{-1} \sum_{i=1}^n \frac{\partial e^{\tilde{\lambda}_n' g_i(\tilde{\beta}_n)}}{\partial \psi} \frac{\partial e^{\tilde{\lambda}_n' g_i(\tilde{\beta}_n)}}{\partial \psi'} \right) \\
&\quad \times (\hat{\psi}_n(\hat{\beta}_n) - \psi_n^*(\beta_n^*)) \\
&= o_p(1), \qquad \text{(C.16)}
\end{aligned}
$$

where the inequality holds by the inequality, $(a+b)^2 \leq 2a^2 + 2b^2$ and the equality holds by $n\omega_n^2 \to \infty$, (B.10), (C.10) and Lemma 5.

The term $W_{n,2}$ is $o_p(1)$ because $0 \leq |W_{n,2}| \leq 2[W_{n,1} \cdot W_{n,3}]^{1/2}$ by the Cauchy–Schwarz inequality. □

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.01.004.

References

Andrews, Donald W.K., 1988. Laws of large numbers for dependent non-identically distributed random variables. Econometric Theory 4, 458–467.

Andrews, Donald W.K., 1992. Generic uniform convergence. Econometric Theory 8, 241–257.

Andrews, Donald W.K., 1994. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. pp. 2247–2294 (Chapter 37).

Andrews, Donald W.K., 1999. Estimation when a parameter is on a boundary. Econometrica 67, 1341–1383.

Andrews, Donald W.K., Soares, Gustavo, 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.

Andrews, Donald W.K., Barwick, Panle Jia, 2012. Inference for parameters defined by moment inequalities: a recommended moment selection procedure. Econometrica 80, 2805–2826.

Andrews, Donald W.K., Guggenberger, Patrik, 2009. Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. Econometric Theory 25, 669–709.

Andrews, Donald W.K., Shi, Xiaoxia, 2013a. Inference based on conditional moment inequality models. Econometrica 81, 609–666.

Andrews, Donald W.K., Shi, Xiaoxia, 2013b. Supplement to 'inference based on conditional moment inequalities'. Econometrica 81, 609–666.

Andrews, Donald W.K., Berry, Steven, Jia, Panle, 2004. Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location. Department of Economics, Yale University.

Berry, Steve, Tamer, Elie, 2006. Identification in models of oligopoly entry. In: Blundel, Newey, Persson (Eds.), Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. 2. pp. 46–85.

Bugni, F.A., 2010. Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. Econometrica 78, 735–753.

Canay, I.A., 2010. El inference for partially identified models: Large deviations optimality and bootstrap validity. J. Econometrics 156, 408–425.

Chen, Xiaohong, Hong, Han, Shum, Matthew, 2007. Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. J. Econometrics 141, 109–140.

Chernozhukov, Victor, Hong, Han, Tamer, Elie, 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75, 1243–1284.

Ciliberto, Federico, Tamer, Elie, 2009. Market structure and multiple equilibria in the airline industry. Econometrica 77, 1791–1828.

Cox, D.R., 1961. Tests of separate families of hypotheses. In: Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability. University of California Press, Berkeley.

Csiszár, I., 1975. I-divergence geometry of probability distributions and minimization problems. Ann. Probab. 3, 146–158.

Gourieroux, Christian, Monfort, Alain, 1994. Testing non-nested hypotheses. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. pp. 2583–2637 (Chapter 44).

Hsu, Yu-Chin, Shi, Xiaoxia, 2013. Model Selection Test for Conditional Moment Inequality Models. Department of Economics, University of Wisconsin at Madison.

Kitamura, Yuichi, 2000. Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood, November. Department of Economics, University of Pennsylvania.

Kitamura, Yuichi, Stutzer, Michael, 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65, 861–874.

Manski, Charles F., 2005. Partial identification with missing data: Concepts and findings. Int. J. Approx. Reason. 39, 151–165.

Manski, Charles F., Tamer, Elie, 2002. Inference on regressions with interval data on a regressor or outcome. Econometrica 70, 519–546.

Mizon, Grayham E., Richard, Jean-Francois, 1986. The encompassing principle and its application to testing non-nested hypotheses. Econometrica 54, 657–678.

Moon, Hyungsik Roger, Schorfheide, Frank, 2009. Estimation with overidentifying inequality moment conditions. J. Econometrics 153, 136–154.

Pakes, Arial, 2010. Alternative models for moment inequalities. Econometrica 78, 1783–1822.

Pakes, Ariel, Porter, J., Ho, Kate, Ishii, Joy, 2007. Moment inequalities and their application. CeMMAP Working Paper No. CWP16/07. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Pesaran, M.H., Weeks, M., 1999. Non-nested hypothesis testing: an overview. Cambridge Working Paper in Economics No. 9918. Faculty of Economics, University of Cambridge.

Ramalho, Joaquim J.S., Smith, Richard J., 2002. Generalized empirical likelihood non-nested tests. J. Econometrics 107, 99–125.

Rivers, Douglas, Vuong, Quang, 2002. Model selection tests for nonlinear dynamic models. Econom. J. 5, 1–39.

Romano, J.P., Shaikh, A.M., 2010. Inference for the identified set in partially identified econometric models. Econometrica 78, 169–211.

Santos, Andres, 2011. Instrumental variable methods for recovering continuous linear functionals. J. Econometrics 161, 129–146.

Shi, Xiaoxia, 2015. A nondegenerate vuong test. Quant. Econ. (forthcoming).

Tamer, Elie, 2003. Incomplete simultaneous discrete response model with multiple equilibria. Rev. Econom. Stud. 70, 147–165.

Vitorino, Maria A., 2012. Empirical entry games with complementarities: An application to the shopping center industry. J. Marketing Res. 49, 175–191.

Vuong, Quang H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.