Optimal Comparison of Misspecified Moment Restriction Models[1]

Vadim Marmer[2] University of British Columbia

Taisuke Otsu[3] Yale University

October 12, 2008

[1]We thank Bruce Hansen, Hiroyuki Kasahara, and Katsumi Shimotsu for helpful comments. Our research is supported by the Social Science and Research Council of Canada under grant number 410-2007-1998 (Marmer) and the National Science Foundation under SES-0720961 (Otsu).

[2]Department of Economics, University of British Columbia, 997 - 1873 East Mall, Vancouver, BC, V6T 1Z1, Canada. Tel.: 1-604-822-8217, email: vadim.marmer@ubc.ca.

[3]Cowles Foundation and Department of Economics, Yale University, P.O. Box 208281, New Haven, CT, 06520-8281, USA. Tel.: 1-203-432-9778, email: taisuke.otsu@yale.edu.

Abstract

This paper considers optimal testing of model comparison hypotheses for misspecified unconditional moment restriction models. We adopt the generalized Neyman-Pearson optimality criterion, which focuses on the convergence rates of the type I and II error probabilities under fixed global alternatives, and derive an optimal but practically infeasible test. We then propose feasible approximation test statistics to the optimal one. For linear instrumental variable regression models, the conventional empirical likelihood ratio test statistic emerges. For general nonlinear moment restrictions, we propose a new test statistic based on an iterative algorithm. We derive the asymptotic properties of these test statistics.

JEL classification: C12; C14; C52
Keywords: Moment restriction; Model comparison; Misspecification; Generalized Neyman-Pearson optimality; Empirical likelihood; GMM

1 Introduction

Econometric models are often defined in the form of moment restrictions and estimated by the generalized method of moments (GMM) (Hansen, 1982), empirical likelihood (EL) (Qin and Lawless, 1994), or their generalizations (see, e.g., Newey and Smith (2004)). In many applications, it is natural to assume that the model is misspecified. While such a model will be rejected with probability approaching one by a consistent overidentifying restriction test, it nevertheless can be of interest as an approximation to the true data generating process. For example, Prescott (1991) argues that a model is only an approximation and should not be regarded as a null hypothesis to be statistically tested. Thus, choosing a model closest to the truth in some sense among several misspecified models is of great importance for practitioners.

Misspecified models and inference procedures for such models have been discussed extensively in the econometric literature. White (1982) studies the properties of the maximum likelihood estimator under misspecification. Vuong (1989) proposes a test of the null hypothesis that two misspecified parametric models provide an equivalent approximation to the true distribution in terms of their Kullback-Leibler information criteria (KLIC). Rivers and Vuong (2002) extend such tests to a more general setting that allows to compare misspecified moment restriction models. Kitamura (2000) develops an information theoretic test that compares misspecified moment restriction models by closeness to the true distribution in terms of the KLIC. Kitamura (2003) extends the information theoretic approach to compare misspecified conditional moment restriction models. Corradi and Swanson (2007) propose a Kolmogorov-type test to compare misspecified dynamic stochastic general equilibrium models. Hall and Inoue (2003) discuss inference for misspecified moment restriction models estimated by the GMM.

This paper considers optimal testing of model comparison hypotheses for misspecified unconditional moment restriction models. Our focus is not on the choice of the measure of fit to set up the model comparison hypotheses but on the choice of the test given the measure of fit. We adopt a GMM-type distance between a model and the true data generating process. To compare different tests, we employ the large deviation approach.[1] In particular, we adopt the generalized Neyman-Pearson (GNP) optimality criterion, which focuses on the convergence rates of the type I and II error

[1] See, e.g., Dembo and Zeitouni (1998) for a review on large deviation theory.

probabilities under fixed global alternatives, and derive an optimal test. The large deviation approach is a natural choice for evaluating the efficiency of an testing procedure when the models are globally misspecified and there is an essential difficulty with formulating Pitman-type local alternatives.

Based on Hoeffding's (1965) seminal work on the large deviation optimality for testing multinomial models, Zeitouni and Gutman (1991), ZG91 hereafter, develop the notion of the GNP optimality and apply it to parameter hypothesis testing problems. Kitamura (2001), K01 hereafter, shows that the EL test is GNP δ-optimal for testing overidentification restrictions. This paper extends their GNP optimality approach to model comparison testing problems. Based on a modified version of the GNP optimality criterion, we derive an optimal test that is defined by the KLIC. However, since the derived optimal test is generally infeasible, we consider approximate tests for specific cases. The approximate test turns out to be the empirical likelihood ratio (ELR) type test (see, for example, equation (4) of K01 for a definition of ELR).

We first consider the model comparison test for linear instrumental variable regression models under the GMM-type distance. In this case, we have an explicit form for the pseudo-true values of the parameters in each model. The approximate GNP optimal test statistic is obtained by solving an EL problem with a constraint given by a smooth function of means. An application of the conventional EL theory (Hall and La Scala, 1990) implies the asymptotic properties of our test.

We next consider the model comparison test for general nonlinear moment restriction models under the GMM-type distance. In this case, our approximate test statistic is defined as a solution to an EL maximization problem subject to a *nonlinear in the multinomial probabilities* constraint, for which an asymptotic theory has not been developed yet. Furthermore, in practice, solving such a problem can be extremely difficult: for a set of multinomial probabilities, one has to perform numerical optimization to obtain the parameters' values, verify the constraint, and then select the set of multinomial probabilities among those that satisfy the constraint. To overcome the practical difficulty, following Wood, Do, and Broom (1996), WDB96 hereafter, we propose an iterative algorithm which requires solving only a sequence of standard *linear in probabilities* EL problems. We also derive the asymptotic properties of the resulted iterated test statistic. Moreover, when the algorithm converges, it converges to the ELR statistic corresponding to the original nonlinear in probabilities problem.

In their paper, WDB96 provide high level assumptions under which the iterated

statistic is asymptotically equivalent to the original ELR statistic. However, the asymptotic distribution of the original ELR statistic in a problem with a nonlinear in probabilities constraint has not been established. On the other hand, we directly derive the asymptotic distribution of the iterated statistic from primitive assumptions.

The problem of comparison of misspecified models should be discerned from non-nested hypothesis testing problems (Davidson and MacKinnon, 1981; MacKinnon, 1983; Smith, 1992). EL-based non-nested tests for moment restriction models are considered by Smith (1997), Ramalho and Smith (2002), and Otsu and Whang (2008). Suppose that the two alternative models are non-nested and therefore cannot be both true at the same time. According to our model comparison null hypothesis, the models have equal measures of fit and, consequently, the null hypothesis implies that they are both misspecified. However, in the literature on non-nested hypothesis testing, the null hypothesis is that one of the models is true. Thus, the two approaches, the non-nested testing and the model comparison testing of misspecified models in the spirit of Vuong (1989), are not competing but rather complementary. The first approach can be used in a search for the true specification, while the later approach can be adopted when the econometrician believes that all alternative models are misspecified or when they all have been rejected by the overidentified restrictions or non-nested tests.

The rest of the paper is organized as follows. Section 2 describes our setup and defines basic concepts. Section 3 discusses the GNP optimality. Section 4 considers linear instrumental variable regression models and derives an approximate optimal test. Section 5 considers general nonlinear moment restriction models and derives a new approximate optimal test. Section 6 concludes. All proofs are given in the Appendix.

We use the following notation. Let E_μ be the expectation under a probability measure μ, $\Pr\{A : \mu\}$ be the probability of an event A under a probability measure μ, I_a be the $a \times a$ unit matrix, $\|z\| = \sqrt{tr(zz')}$ be the Euclidean norm for a vector or matrix z, $\|z\|_W = \sqrt{z'Wz}$ for a vector z and a symmetric positive definite matrix W, and $cl(A)$ and $int(A)$ be the closure and interior of a set A, respectively.

2 Setup and Definitions

Suppose that we observe an iid sample $\{w_i\}_{i=1}^n$ that is drawn from the true and unknown distribution law μ_0 for the random vector w with the support in $\mathbb{R}^q$. Consider the unconditional moment restriction model implied by some economic theory:

$$E_{\mu_0} g(w, \theta_0) = 0, \tag{1}$$

where $g : \mathbb{R}^q \times \Theta \to \mathbb{R}^{l_g}$ is a known function up to the unknown parameters $\theta_0 \in \Theta \subset \mathbb{R}^{p_g}$ with $l_g > p_g$ (overidentified). In this paper, we denote moment restriction models by their corresponding moment functions. For example, the model (1) is called the model g. If the model g is correctly specified, i.e., (1) is satisfied at some parameter value of θ_0, then we can apply the standard GMM theory for estimation and inference on θ_0.

The focus of this paper is to compare two (or more) misspecified moment restriction models. To formalize our idea, we introduce some notation. Let $\mathcal{M}$ be the space of all probability measures on $\mathbb{R}^q$ and define

$$\mathcal{P}_\theta^g = \{\mu \in \mathcal{M} : E_\mu g(w, \theta) = 0\}, \quad \mathcal{P}^g = \cup_{\theta \in \Theta} \mathcal{P}_\theta^g,$$

i.e., $\mathcal{P}_\theta^g$ is a set of measures satisfying the moment restriction by g at θ, and $\mathcal{P}^g$ is a set of measures satisfying the moment restriction at some $\theta \in \Theta$. Then misspecification of the model (1) is defined as follows.

Definition 1 (Misspecification) *The model g is said to be misspecified if $\mu_0 \notin \mathcal{P}^g$.*

Remark. By Definition 1 and because $l_g > p_g$, it follows that $\inf_{\theta \in \Theta} \|E_{\mu_0} g(w, \theta)\| > 0$, if g is misspecified.

The alternative moment restriction model is defined similarly to g:

$$E_{\mu_0} h(w, \beta_0) = 0, \tag{2}$$

where $h : \mathbb{R}^q \times B \to \mathbb{R}^{l_h}$ is another moment restriction function with $B \subset \mathbb{R}^{p_h}$ and $l_h > p_h$. For the model h, we also define the sets of measures $\mathcal{P}_\beta^h = \{\mu \in \mathcal{M} : E_\mu h(\beta) = 0\}$ and $\mathcal{P}^h = \cup_{\beta \in B} \mathcal{P}_\beta^h$.

We consider the situation where the both models g and h are misspecified, and we want to compare these misspecified models in terms of their closeness to the true measure μ_0. Let $D(g, \mu_0)$ (or $D(h, \mu_0)$) be the distance between the model g (or h) and the true measure μ_0. For example, Vuong (1989) and Kitamura (2000) adopt the KLIC to define the distance. The KLIC-based distance D_{KLIC} is defined as

$$D_{KLIC}(g, \mu_0) = \inf_{\mu \in \mathcal{P}^g} I(\mu_0 \| \mu), \quad D_{KLIC}(h, \mu_0) = \inf_{\mu \in \mathcal{P}^h} I(\mu_0 \| \mu),$$

where $I(\mu_0 \| \mu)$ is the KLIC from μ_0 to μ:

$$I(\mu_0 \| \mu) = \begin{cases} \int \log\left(\frac{d\mu_0}{d\mu}\right) d\mu_0, & \text{if } \mu_0 \ll \mu, \\ \infty, & \text{otherwise.} \end{cases}$$

Given a distance D, based on Vuong (1989) and Rivers and Vuong (2002), the model comparison test is defined as follows.

Definition 2 (Model comparison test) *The model comparison test between the models g and h under the distance D is to test the null hypothesis*

$$H_0 : D(g, \mu_0) = D(h, \mu_0), \tag{3}$$

against the alternative hypothesis $H_1 = \{H_g \text{ or } H_h\}$, where

$$H_g : D(g, \mu_0) < D(h, \mu_0) \quad (g \text{ is preferred over } h),$$
$$H_h : D(g, \mu_0) > D(h, \mu_0) \quad (h \text{ is preferred over } g).$$

When all competing models are misspecified, their comparison and relative ranking depend crucially on the choice of the distance: different definitions of the distance can lead to different ranking of the models. In this paper (in Sections 4 and 5) we focus on the GMM-type distance. For given symmetric and positive definite matrices W_g and W_h, the GMM-type distance between the model g and the true distribution μ_0 (and h and μ_0) is defined as

$$D_{GMM}(g, \mu_0) = \inf_{\theta \in \Theta} \|E_{\mu_0} g(w, \theta)\|^2_{W_g}, \quad D_{GMM}(h, \mu_0) = \inf_{\beta \in B} \|E_{\mu_0} h(w, \beta)\|^2_{W_h}. \tag{4}$$

We focus on the GMM-type distance rather than the KLIC because D_{GMM} directly

punishes for the magnitude of the violation of the moment condition. In the case of D_{KLIC}, the model h is preferred over g if the family of measures $\mathcal{P}^h$ is closer to μ_0 than $\mathcal{P}^g$ in terms of KLIC. However, the magnitude of violation of the moment restrictions is not considered directly when the models are compared. We believe that comparison in terms of D_{GMM} is more attractive than comparison in terms of the KLIC distance when $\inf_{\theta\in\Theta} \|E_{\mu_0} g(w,\theta)\|_{W_g}$ is meaningful from the economic theory perspective.

Once the distance is chosen and the null hypothesis is defined, one can address the issue of optimal testing for the selected distance D. In the case of D_{GMM}, the null hypothesis can be tested by using the difference between the sample analogues of $\inf_{\theta\in\Theta} \|E_{\mu_0} g(w,\theta)\|^2_{W_g}$ and $\inf_{\beta\in B} \|E_{\mu_0} h(w,\beta)\|^2_{W_h}$:

$$\inf_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta) \right\|^2_{W_g} - \inf_{\beta\in B} \left\| \frac{1}{n}\sum_{i=1}^{n} h(w_i,\beta) \right\|^2_{W_h}. \tag{5}$$

This approach was considered by Rivers and Vuong (2002). In this paper, we develop an alternative testing procedure motivated by the notion of the GNP optimality. The GNP optimality and its corresponding testing rule are discussed in the next section.

3 GNP Optimal Test

In this section, we investigate optimality for the model comparison test in Definition 2. Our optimality criterion is based on the global properties test statistics, in particular the behaviors of error probabilities under fixed alternatives. If a test is consistent, the type I and II error probabilities of the test typically decrease to zero at an exponential rate under fixed alternatives, and competing tests can be compared by the convergence rates of their error probabilities.

An alternative way to evaluate test statistics is to compute their local power functions under a sequence of (Pitman-type) drifting local alternatives, which converges to some measure satisfying H_0 in (3). However, since we are interested in the case where both models are globally misspecified, it is natural to investigate the global behaviors of tests under the fixed true measure.

Among several optimality criteria based on global properties of tests (see, for example, Chapter 12 of Serfling (1980)), we adopt the GNP optimality criterion

developed by ZG91 and K01, among others. Let μ_n be the empirical measure based on the sample $\{w_i\}_{i=1}^n$:

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(w_i),$$

for a set A in the Borel σ-field on R^q, where 1_A is the indicator function of A, and define

$$\mathcal{P}_0 = \{\mu \in \mathcal{M} : D(g,\mu) = D(h,\mu)\}.$$

Note that $\mathcal{P}_0$ is the set of measures satisfying the null hypothesis H_0. Consider a test $\Omega = (\Omega_0, \Omega_1)$ based on μ_n defined by the partition (Ω_0, Ω_1) for $\mathcal{M}$, that is[2]

$$\begin{aligned} &\text{accept } H_0 \text{ if } \ \mu_n \in \Omega_0, \\ &\text{reject } H_1 \text{ if } \mu_n \in \Omega_1 = \mathcal{M} \setminus \Omega_0. \end{aligned}$$

Then the type I and II error probabilities are defined as

$$\begin{aligned} &\Pr\{\mu_n \in \Omega_1 : \mu_0\} \quad \text{for} \ \ \mu_0 \in \mathcal{P}_0, \\ &\Pr\{\mu_n \in \Omega_0 : \mu_0\} \quad \text{for} \ \ \mu_0 \notin \mathcal{P}_0, \end{aligned}$$

respectively.

The original idea of the Neyman-Pearson optimality is to minimize the type II error probability under a restriction on the type I error probability in finite samples. However, since it is generally difficult to establish this original Neyman-Pearson optimality, we commonly focus on the large sample properties of the test. If the test is consistent, both error probabilities converge to zero under fixed alternatives and their convergence rates are typically exponential. By modifying the idea of the original Neyman-Pearson optimality to the convergence rate analogs, the GNP optimality criterion is described as

$$\begin{aligned} &\text{minimize } \lim_{n\to\infty} \tfrac{1}{n}\log \Pr\{\mu_n \in \Omega_0 : P_1\} \quad \text{for each } P_1 \in \mathcal{M} \setminus \mathcal{P}_0, \\ &\text{subject to } \sup_{P_0 \in \mathcal{P}_0} \lim_{n\to\infty} \tfrac{1}{n}\log \Pr\{\mu_n \in \Omega_1 : P_0\} \le -\alpha. \end{aligned} \tag{6}$$

[2]We focus on the class of tests defined by a partition for the empirical measure. We conjecture that an analogous argument to Lemma 1 of ZG91 may yield a sufficiency result to restrict on this class of tests. For example, the usual GMM-type test statistic in (5) can be written as $D_{GMM}(g,\mu_n) - D_{GMM}(h,\mu_n)$.

To analyze these convergence rates of the error probabilities, we can apply the large deviation theory for the empirical measure. In particular, Sanov's theorem is useful for our purpose.

Lemma 3 (Sanov's Theorem) *Suppose that* $\{w_i\}_{i=1}^n$ *is an iid sample from* $\mu_0 \in \mathcal{M}$*. Then its empirical measure* μ_n *satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr\{\mu_n \in G : \mu_0\} \leq -\inf_{\nu\in G} I(\nu\|\mu_0),$$

for any closed set $G \subset \mathcal{M}$ *with respect to the Lévy metric, and*

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr\{\mu_n \in H : \mu_0\} \geq -\inf_{\nu\in H} I(\nu\|\mu_0),$$

for any open set $H \subset \mathcal{M}$ *with respect to the Lévy metric.*

The proof of Sanov's theorem can be found in Deuschel and Stroock (1989), for example. Sanov's theorem says that the error probabilities written in terms of the empirical measure are determined by the KLIC between the data generating measure μ_0 and the sets of interest G and H. This result is particularly useful for establishing the bounds on the convergence rates of the type I and II errors probabilities. It also suggests that a test based on the KLIC distance between the set of measures satisfying H_0 and the empirical measure might enjoy the GNP optimal property. On the other hand, Sanov's theorem has some rough nature: we can only obtain the upper (or lower) bound for closed (or open) sets with respect to the Lévy metric. In general, however, the rejection regions defined in terms of the KLIC do not have to be closed, which makes derivation of the GNP optimality in the sense of (6) very difficult (see ZG91 and K01 for more discussions). Therefore, we consider the following modified version of the GNP optimality called the GNP δ-optimality.

Let $D_L(\mu,\nu)$ denote the Lévy metric between $\mu \in \mathcal{M}$ and $\nu \in \mathcal{M}$:

$$D_L(\mu,\nu) = \inf\{\epsilon > 0 : F_\mu(w - \epsilon\iota) - \epsilon \leq F_\nu(w) \leq F_\mu(w - \epsilon\iota) + \epsilon \quad \text{for all } w \in \mathbb{R}^q\},$$

where F_μ and F_ν are the distribution functions of μ and ν, respectively, and $\iota = (1,\ldots,1)' \in \mathbb{R}^q$. Let $B(\mu,\delta) = \{\nu \in \mathcal{M} : D_L(\mu,\nu) < \delta\}$ be an open ball around $\mu \in \mathcal{M}$ with radius $\delta > 0$. For a test $\Omega = (\Omega_0, \Omega_1)$, define the partition $\Omega^\delta = \left(\Omega_0^\delta, \Omega_1^\delta\right)$

with $\Omega_1^\delta = \cup_{\mu\in\Omega_1} B(\mu,\delta)$ and $\Omega_0^\delta = \mathcal{M}\setminus\Omega_1^\delta$. The set Ω_1^δ is often called the δ-blowup (or δ-smoothing) of the critical region Ω_1 by the Lévy ball.

Definition 4 (GNP δ-optimality) *A test defined by the partition $\Lambda = (\Lambda_0, \Lambda_1)$, which may depend on δ, is called GNP δ-optimal if for each $\delta > 0$,*

(a) $\sup_{P_0\in\mathcal{P}_0} \limsup_{n\to\infty} \frac{1}{n}\log \Pr\left\{\mu_n \in \Lambda_1^\delta : P_0\right\} \le -\alpha$ *for some $\alpha > 0$,*

(b) *for any test $\Omega = (\Omega_0, \Omega_1)$ satisfying*

$$\sup_{P_0\in\mathcal{P}_0} \limsup_{n\to\infty} \frac{1}{n}\log \Pr\left\{\mu_n \in \Omega_1^{\bar\delta} : P_0\right\} \le -\alpha \quad \text{for some } \bar\delta > \delta,$$

we have

$$\limsup_{n\to\infty} \frac{1}{n}\log \Pr\left\{\mu_n \in \mathcal{M}\setminus\Lambda_1^\delta : P_1\right\} \le \limsup_{n\to\infty} \frac{1}{n}\log \Pr\left\{\mu_n \in \mathcal{M}\setminus\Omega_1^\delta : P_1\right\},$$

for all $P_1 \in \mathcal{M}\setminus\mathcal{P}_0$.

Based on ZG91, we consider the following (δ dependent) KLIC-based test $\Lambda_\delta = (\Lambda_{0,\delta}, \Lambda_{1,\delta})$:

$$\begin{aligned}
\text{accept } H_0 \text{ if } \mu_n \in \Lambda_{0,\delta} &= \left\{\nu\in\mathcal{M} : \inf_{\mu\in\mathcal{P}_0} \inf_{\nu'\in cl(B(\nu,c\delta))} I(\nu'\|\mu) \le \alpha\right\},\\
\text{reject } H_0 \text{ if } \mu_n \in \Lambda_{1,\delta} &= \mathcal{M}\setminus\Lambda_{0,\delta},
\end{aligned}$$

for some $c > 1$. In other words, we test H_0 by the test statistic

$$T_{n,\delta} = \inf_{\mu\in\mathcal{P}_0} \inf_{\nu\in cl(B(\mu_n,c\delta))} I(\nu\|\mu), \tag{7}$$

with the critical value α. The following theorem provides the GNP δ-optimality of the KLIC-based test Λ_δ.

Theorem 5 (GNP δ-optimal test) *Suppose that $\{w_i\}_{i=1}^n$ is iid and the set $\{\nu\in\mathcal{M} : \inf_{\mu\in\mathcal{P}_0} I(\nu\|\mu) \le \alpha\}$ is compact with respect to the Lévy metric. Then the KLIC-based test Λ_δ is GNP δ-optimal to test the model comparison hypothesis H_0 against H_1.*

Remarks. **(a)** The iid assumption on the data $\{w_i\}_{i=1}^n$ is required to apply Sanov's theorem. Under weakly dependent data, large deviation properties of the empirical measure can be analyzed by the Gärtner-Ellis theorem (Dembo and Zeitouni, 1998, Theorem 2.3.6), where the convergence rate is characterized by the long-run limit of the moment generating function instead of the KLIC. Although it is beyond the scope of this paper, we conjecture that a test statistic based on this rate function can yield an analogous optimality result.

(b) To prove the GNP optimality of a test in the sense of (6) by Sanov's theorem, one needs closedness of the rejection region $\{\nu \in \mathcal{M} : \inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu) \geq \alpha\}$ which is generally not true (see the discussion on page 287, Section III in ZG91). For GNP δ-optimality, we impose a weaker condition that the set $\{\nu \in \mathcal{M} : \inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu) \leq \alpha\}$ is compact. The later condition holds if, for example, if $\inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu)$ is lower semicontinuous in ν under the Lévy metric.

(c) The compactness assumption on the set $\{\nu \in \mathcal{M} : \inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu) \leq \alpha\}$ restricts the form of the null hypothesis $\mathcal{P}_0$, i.e., not only the forms of the moment functions g and h, but also the form of the distance D. For example, suppose that the distances $D(g, \mu_0)$ and $D(h, \mu_0)$ are continuous in μ_0 under the Lévy metric, which is satisfied if g and h are bounded and the GMM-type distance D_{GMM} in (4) is adopted. In this case, applications of the maximum theorem (Leininger, 1984) combined with the lower semicontinuity of the KLIC (Chaganty and Karandikar, 1996) imply the lower semicontinuity of $\inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu)$ in ν under the Lévy metric, which in turn implies the compactness of $\{\nu \in \mathcal{M} : \inf_{\mu \in \mathcal{P}_0} I(\nu \| \mu) \leq \alpha\}$ under the Lévy metric. Note that the KLIC distance D_{KLIC} does not guarantee the continuity of $D_{KLIC}(g, \mu_0)$ and $D_{KLIC}(h, \mu_0)$ in μ_0 in general even if g and h are bounded.

(d) Although the GNP δ-optimality can be considered as an weaker definition of optimality than the original Neyman-Pearson or the GNP optimality in the sense of (6), this theorem is insightful: we should take the minimum distance between the null space $\mathcal{P}_0$ and the closed Lévy ball $cl(B(\mu_n, 2\delta))$ around the empirical measure μ_n by using the KLIC.

(e) The obvious limitation of this theorem is the fact that both the optimal test Λ^δ and alternative test Ω^δ depend on the blowup constant δ. For the optimal test Λ^δ, we can apply a similar argument to Corollary 3 of ZG91 and construct a positive and monotone decreasing sequence $\{\delta_n\}_{n \in \mathbb{N}}$ with $\delta_n \to 0$ such that the n-dependent test $\left\{\Lambda^{\delta_n}\right\}_{n \in \mathbb{N}}$ satisfies the GNP δ-optimality. On the other hand, for the alternative

test Ω^δ, suppose that the test Ω^δ is "regular" in the sense of ZG91, i.e.

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in\Omega_1^\delta : P_0\right\}=\limsup_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in\Omega_1 : P_0\right\},$$

for each $P_0\in\mathcal{P}_0$, which is satisfied when $\inf_{\mu\in int(\Omega_1)} I(\mu\|P_0)=\inf_{\mu\in cl(\Omega_1)} I(\mu\|P_0)$ (see Lemma 4 of ZG91). Then we can replace the blowup critical regions $\Omega_1^{\bar{\delta}}$ and Ω_1^δ in Definition 4 with the original one Ω_1.

(f) It is generally difficult to compute the test statistic $T_{n,\delta}$ in practice. However, inspired by this optimality result, we can consider feasible approximations to the GNP δ-optimal test $T_{n,\delta}$. In particular, we remove the δ-blowup and focus on the approximate test statistic

$$T_n=\inf_{\mu\in\mathcal{P}_0} I(\mu_n\|\mu). \tag{8}$$

(g) When μ_0 is discrete, the approximate statistic T_n is GNP δ-optimal and no smoothing is required (see Section II in ZG91). The GNP optimality of the likelihood ratio test in multinomial models is established by Hoeffding (1965).

(h) The approximate statistic T_n is actually equivalent to the ELR statistic. Let p be a discrete measure that assigns strictly positive probabilities to the observations $\{w_i\}_{i=1}^n$, i.e. $p(A)=\sum_{i=1}^n p_i 1_A(w_i)$, where $p_i>0$ is the probability assigned by p to w_i. The ELR statistic is defined as

$$ELR_n=-2\log\left(L_{EL}^{constrained}/L_{EL}^{unconstrained}\right),$$

where $L_{EL}^{constrained}$ is the constrained EL:

$$\begin{aligned} L_{EL}^{constrained}= \ & \max_{p_1,\ldots,p_n}\prod_{i=1}^n p_i, \\ & \text{s.t. } p_i>0, \textstyle\sum_{i=1}^n p_i=1, D(g,p)=D(h,p) \end{aligned} \tag{9}$$

and $L_{EL}^{unconstrained}$ is the maximum of EL without the constraint $D(g,p)=D(h,p)$:

$$L_{EL}^{unconstrained}=n^{-n}$$

(when there is no p that satisfies the constraint, set $ELR_n=\infty$). Let p^* be a solution to the maximization problem in (9). Due to the definition of the KLIC, one only has

to consider the discrete measures p when solving (8). Now,

$$\begin{aligned} T_n &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{1/n}{p_i^*} \\ &= \frac{1}{2n}ELR_n. \end{aligned}$$

The ELR statistic has an asymptotic χ^2 null distribution when the constraint is linear in probabilities (Owen, 2001) or can be expressed as a smooth function of means (Hall and La Scala, 1990). In our case, the constraint is generally nonlinear in probabilities.

The next two sections investigate implementation and the statistical properties of the approximate optimal test based on T_n.

4 Approximate Optimal Test: Linear Case

This section considers a comparison of linear instrumental variable regression models. The linear case is particularly illustrating since the testing problem can be expressed in a closed form in terms of smooth functions of means, and the asymptotic null distribution of the test statistic can be obtained directly from the existing results in the EL literature.

Let $w = \left(y^g, x^{g\prime}, z^{g\prime}, y^h, x^{h\prime}, z^{h\prime}\right)'$, where y^g, x^g, and z^g denote the dependent variable, endogenous regressors, and instruments in the model g; and y^h, x^h, and z^h denote the dependent variable, endogenous regressors, and instruments in the model h. There can be overlapping variables between the two models; for example, it is possible that the two models have the same dependent variables ($y^g = y^h = y$) and regressors ($x^g = x^h = x$) but different sets of instruments. The moment restrictions are:

$$E_{\mu_0} g\left(w, \theta_0\right) = E_{\mu_0} z^g\left(y^g - x^{g\prime}\theta_0\right) = 0, \tag{10}$$

$$E_{\mu_0} h\left(w, \beta_0\right) = E_{\mu_0} z^h\left(y^h - x^{h\prime}\beta_0\right) = 0, \tag{11}$$

where $\theta_0 \in \Theta \subset \mathbb{R}^{p_g}$ and $\beta_0 \in B \subset \mathbb{R}^{p_h}$, and it is assumed that for no value of the parameters the moment conditions (10) and (11) are satisfied. We consider the model comparison test under the GMM-type distance D_{GMM} in (4), i.e., the null hypothesis is

$$H_0 : \inf_{\theta\in\Theta} \left\| E_{\mu_0} z^g \left(y^g - x^{g\prime}\theta \right) \right\|^2_{W_g} = \inf_{\beta\in B} \left\| E_{\mu_0} z^h \left(y^h - x^{h\prime}\beta \right) \right\|^2_{W_h}. \tag{12}$$

An important condition is that the models are overidentified, otherwise, in the exactly identified case, the null restriction is trivially satisfied with zeros on both sides.

Assumption 6 **(a)** *The models are overidentified:* $rank\left(E_{\mu_0} z^g x^{g\prime}\right) = l_g > p_g$ *and* $rank\left(E_{\mu_0} z^h x^{h\prime}\right) = l_g > p_g$.

(b) *The models* g *and* h *are misspecified.*

By part (a) of the assumption, we have closed-form solutions for the minimization problems in (12):

$$\theta^*\left(\mu_0\right) = \left(\left(E_{\mu_0} x^g z^{g\prime}\right) W_g \left(E_{\mu_0} z^g x^{g\prime}\right)\right)^{-1} \left(E_{\mu_0} x^g z^{g\prime}\right) W_g \left(E_{\mu_0} z^g y^g\right),$$

with a similar expression for $\beta^*\left(\mu_0\right)$. Thus, the null hypothesis in (12) can be written as a function of the means:

$$H_0 : f\left(\eta\left(\mu_0\right)\right) = 0,$$

where

$$\begin{aligned} \eta\left(\mu_0\right) &= \left(E_{\mu_0} z^g x^{g\prime}, E_{\mu_0} z^g y^g, E_{\mu_0} z^h x^{h\prime}, E_{\mu_0} z^h y^h\right), \\ f\left(\eta\left(\mu_0\right)\right) &= \left\| E_{\mu_0} z^g y^g - \left(E_{\mu_0} z^g x^{g\prime}\right) \theta^*\left(\mu_0\right) \right\|^2_{W_g} - \left\| E_{\mu_0} z^h y^h - \left(E_{\mu_0} z^h x^{h\prime}\right) \beta^*\left(\mu_0\right) \right\|^2_{W_h}. \end{aligned}$$

In this case, our approximate GNP δ-optimal test statistic in (8) takes the following form:

$$\begin{aligned} T_n^L &= \inf_{\{\mu\in\mathcal{M}: f(\eta(\mu))=0\}} I\left(\mu_n \| \mu\right) \\ &= \min_{\{\eta: f(\eta)=0\}} \ell\left(\eta\right), \end{aligned} \tag{13}$$

where $\ell\left(\eta\right) = \ell\left(\eta_1^g, \eta_2^g, \eta_1^h, \eta_2^h\right)$ and

$$\begin{aligned} \ell\left(\eta_1^g, \eta_2^g, \eta_1^h, \eta_2^h\right) = - \ & \max_{\{p_i\}_{i=1}^n} \ \frac{1}{n}\sum_{i=1}^n \log\left(np_i\right) \\ & \text{s.t.} \quad p_i > 0, \quad \sum_{i=1}^n p_i = 1, \\ & \quad \sum_{i=1}^n p_i z_i^g x_i^{g\prime} = \eta_1^g, \quad \sum_{i=1}^n p_i z_i^g y_i^g = \eta_2^g, \\ & \quad \sum_{i=1}^n p_i z_i^h x_i^{h\prime} = \eta_1^h, \quad \sum_{i=1}^n p_i z_i^h y_i^h = \eta_2^h. \end{aligned}$$

Note that the derived test statistic T_n^L is equivalent to the ELR statistic for a smooth function of means by Hall and La Scala (1990). Therefore, in this setup, the ELR test has a rationale as an approximation to the GNP δ-optimal test discussed in the last section. To derive the asymptotic property of the test statistic, we can directly apply the existing result by Hall and La Scala (1990). To this end, we impose the following assumption.

Assumption 7 *The vector* $\left(vec\left(z^g x^{g\prime}\right)', \left(z^g y^g\right)', vec\left(z^h x^{h\prime}\right)', \left(z^h y^h\right)'\right)'$ *excluding the overlapping elements has a finite and positive definite variance matrix.*

From Theorem 2.1 of Hall and La Scala (1990), we have the following result:

Theorem 8 (Approximate optimal test for linear models) *Consider the distance* D_{GMM}*, and suppose that Assumptions 6 and 7 hold. Then the model comparison test statistic* T_n^L *between the models (10) and (11) satisfies* $2nT_n^L \to_d \chi_1^2$ *under* H_0*, and* $2nT_n^L \to \infty$ *almost surely under* H_1*.*

5 Approximate Optimal Test: General Case

This section considers general nonlinear moment restriction models (1) and (2). As in the last section, we consider the model comparison test under the GMM-type distance D_{GMM}. However, in contrast to the linear case, the solutions $\theta^*(\mu_0)$ and $\beta^*(\mu_0)$ to $\inf_{\theta\in\Theta} \|E_{\mu_0} g(w,\theta)\|_{W_g}^2$ and $\inf_{\beta\in B} \|E_{\mu_0} h(w,\beta)\|_{W_h}^2$, respectively, cannot be expressed in a closed form.

Hereafter for brevity we also use the notation

$$g_i(\theta) = g(w_i,\theta) \text{ and } h_i(\beta) = h(w_i,\beta).$$

Similarly to (13), the approximate GNP δ-optimal test can be written as

$$\begin{aligned} T_n^G &= \inf_{\mu\in\left\{\mu\in\mathcal{M}:\inf_{\theta\in\Theta}\|E_\mu g(w,\theta)\|_{W_g}^2 = \inf_{\beta\in B}\|E_\mu h(w,\beta)\|_{W_h}^2\right\}} I(\mu_n\|\mu) \\ &= -\max_{\{p_i\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n \log(np_i) \\ &\quad \text{s.t.} \quad p_i > 0, \quad \sum_{i=1}^n p_i = 1, \\ &\qquad \inf_{\theta\in\Theta}\left\|\sum_{i=1}^n p_i g_i(\theta)\right\|_{W_g}^2 = \inf_{\beta\in B}\left\|\sum_{i=1}^n p_i h_i(\beta)\right\|_{W_h}^2. \end{aligned} \tag{14}$$

Note that since the last restriction in the above maximization problem is nonlinear in p_i, we cannot apply the standard implementation and asymptotic theory of EL.

The difficulties of EL in the case of nonlinear in probabilities constraints have been addressed by WDB96. They proposed an iterative algorithm based on the approximate linearization of the constraints. Instead of solving the EL problem with a nonlinear in probabilities constraint, the econometrician solves a sequence of EL problems with linear in probabilities constraints that hold approximately. The linearization is obtained by a Taylor expansion of the original constraint. A similar idea can be used in our case as well. However, unlike WDB96, we do not rely on a Taylor expansion to obtain a linearization, and the linearization is exact in our case in some sense clarified later (see Remark (b) following Lemma 10 below).

Define the Hessians of the GMM objective functions:

$$H_g(\mu,\theta) = \left(E_\mu \frac{\partial g(w,\theta)}{\partial \theta'}\right)' W_g \left(E_\mu \frac{\partial g(w,\theta)}{\partial \theta'}\right) + \\ + \left(I_{p_g} \otimes (W_g E_\mu g(w,\theta))\right)' \left(E_\mu \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g(w,\theta)}{\partial \theta'}\right)\right),$$

and

$$H_h(\mu,\beta) = \left(E_\mu \frac{\partial h(w,\beta)}{\partial \beta'}\right)' W_h \left(E_\mu \frac{\partial h(w,\beta)}{\partial \beta'}\right) + \\ + \left(I_{p_h} \otimes (W_h E_\mu h(w,\beta))\right)' \left(E_\mu \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h(w,\beta)}{\partial \beta'}\right)\right).$$

To characterize the solution to the likelihood maximization problem (14), we make the following assumption.

Assumption 9

(a) *Θ and B are compact.*

(b) *The model g is misspecified and $\|E_{\mu_0} g_i(\theta)\|^2_{W_g}$ has a unique minimum at $\theta^* \in int(\Theta)$; the model h is misspecified and $\|E_{\mu_0} h_i(\beta)\|^2_{W_h}$ has a unique minimum at $\beta^* \in int(B)$.*

(c) *$g_i(\theta)$ is twice continuously differentiable on Θ almost surely; $h_i(\beta)$ is twice continuously differentiable on B almost surely.*

(d) *$H_g(\mu_0, \theta)$ is nonsingular in a neighborhood N_{θ^*} of θ^*; $H_h(\mu_0, \beta)$ is nonsingular in a neighborhood N_{β^*} of β^*.*

Assumptions 9(a) and (c) are standard for nonlinear models. Assumptions 9(b) requires uniqueness of the pseudo-true values, which is often assumed in the literature of misspecification analysis (Vuong, 1989; Kitamura, 2000; Rivers and Vuong, 2002; Hall and Inoue, 2003). Assumptions 9(d) requires that the Hessians of the GMM objective functions are nonsingular. Such an assumption appears, for example, in Hall and Inoue (2003). Note that, in comparison to the standard correctly specified case, the Hessian involves an extra term when the model is misspecified. For example, the second summand in the definition of H_g is different from zero even when H_g is evaluated at μ_0 and θ^*.

Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ be the solution to (14), λ be the Lagrange multiplier associated with the last constraint in (14), $\theta(p) = \arg\inf_{\theta\in\Theta} \|\sum_{i=1}^n p_i g_i(\theta)\|^2_{W_g}$, and $\beta(p) = \arg\inf_{\beta\in B} \|\sum_{i=1}^n p_i h_i(\beta)\|^2_{W_h}$. The following lemma characterizes the solution $\hat{p}$ and the associated parameter values $\hat{\theta} = \theta(\hat{p})$ and $\hat{\beta} = \beta(\hat{p})$ as a solution to an EL maximization problem with a linear moment restriction.

Lemma 10 (Linearization) *Under Assumption 9, $\left(\hat{p}, \hat{\theta}, \hat{\beta}\right)$ satisfies*

$$\left(\hat{\theta}, \hat{\beta}, \hat{\lambda}\right) = \arg \inf_{\theta\in\Theta, \beta\in B} \max_{\lambda\in\mathbb{R}} \sum_{i=1}^n \log\left(1 + \lambda d_i\left(\theta, \beta; \hat{p}, \hat{\theta}, \hat{\beta}\right)\right), \tag{15}$$

where

$$d_i\left(\theta, \beta; \hat{p}, \hat{\theta}, \hat{\beta}\right) = g_i(\theta)' W_g \left(\sum_{j=1}^n \hat{p}_j g_j\left(\hat{\theta}\right)\right) - h_i(\beta)' W_h \left(\sum_{j=1}^n \hat{p}_j h_j\left(\hat{\beta}\right)\right),$$

and

$$\hat{p}_i = \frac{1}{n\left(1 + \hat{\lambda} d_i\left(\hat{\theta}, \hat{\beta}; \hat{p}, \hat{\theta}, \hat{\beta}\right)\right)}.$$

Remarks. (a) Lemma 10 represents $\hat{\theta}$, $\hat{\beta}$, and $\hat{p}$ as a solution to the fixed point problem. Given $\hat{\theta}$, $\hat{\beta}$, and $\hat{p}$, one can construct a linear in probabilities EL problem that has $\hat{\theta}$ and $\hat{\beta}$ as a solution. The probabilities $\hat{p}$ that solve the original EL problem can be also recovered obtained as a by-product of solving the linearized EL problem in (15).

(b) Unlike the case considered in WDB96, here the solution does not depend on the derivatives $\partial\theta(p)/\partial p_i$ and $\partial\beta(p)/\partial p_i$. Our linearization is exact in this sense. This is due to the fact that we have quadratic functions in the constraint in (14).

Since $\hat{\theta}$, $\hat{\beta}$, and $\hat{p}$ are solutions to a fixed point problem, Lemma 10 leads to the following iterative algorithm.

Step 1: Given $\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}, \hat{p}^{(s-1)}\right)$, update the estimators of θ and β by solving the minimax problem:

$$\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}, \hat{\lambda}^{(s)}\right) = \arg \inf_{\theta\in\Theta,\beta\in B} \max_{\lambda\in\mathbb{R}} \sum_{i=1}^{n} \log\left(1 + \lambda d_i\left(\theta, \beta; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right). \tag{16}$$

Step 2: Update $\hat{p}$ using $\hat{\theta}^{(s)}$, $\hat{\beta}^{(s)}$, and $\hat{\lambda}^{(s)}$ obtained at Step 1:

$$\hat{p}_i^{(s)} = \frac{1}{n\left(1 + \hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right)}.$$

Iterate Steps 1 and 2 until convergence of the sequence $\left\{\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}\right) : s \in \mathbb{N}\right\}$ is achieved or the maximum allowed number of iterations is reached.

Remarks. **(a)** At each iteration, we solve a standard linear in probabilities EL maximization problem with the moment function $d_i\left(\theta, \beta; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)$. Since the Lagrangian λ is scalar, it is cheap to implement the maximization in (16).

(b) The iterated test statistic to approximate T_n^G in (14) is obtained as

$$T_n^{G(s)} = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right). \tag{17}$$

(c) We suggest terminating the algorithm by checking a convergence criterion on $\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}\right)$ rather than on the value of the objective function in (16) or $T_n^{G(s)}$ because the constraint changes from iteration to iteration. Thus, in the absence of convergence of the sequence $\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}\right)$, we suggest setting $T_n^G = \infty$.

(d) As we show in the proof of Lemma 10, the original and linearized problems satisfy the same first-order conditions. Therefore, if the algorithm converges, the sequence of the iterated statistics $T_n^{G(s)}$ converges to the original statistic T_n^G.

(e) Similar algorithms motivated by fixed point problems have been considered in the econometric literature in the context of discrete dynamic games (Aguirregabiria and Mira, 2002, 2007; Kasahara and Shimotsu, 2008).

Next, we discuss the asymptotic property of the test statistic obtained from the iterated procedure described above. We add the following assumptions.

Assumption 11

(a) $0 < E_{\mu_0} d_i^2 (\theta^*, \beta^*; \mu_0, \theta^*, \beta^*) < \infty$.

(b) *In a neighborhood* N_{θ^*} *of* θ^* *and for some* $\varepsilon > 0$, $E_{\mu_0} \sup_{\theta \in N_{\theta^*}} \|g_i(\theta)\|^{2+\varepsilon} < \infty$; *in a neighborhood* N_{β^*} *of* β^*, $E_{\mu_0} \sup_{\beta \in N_{\beta^*}} \|h_i(\beta)\|^{2+\varepsilon} < \infty$.

(c) $E_{\mu_0} \sup_{\theta \in N_{\theta^*}} \left\| \frac{\partial g_i(\theta)}{\partial \theta'} \right\|^2 < \infty$; $E_{\mu_0} \sup_{\beta \in N_{\beta^*}} \left\| \frac{\partial h_i(\beta)}{\partial \beta'} \right\|^2 < \infty$.

(d) $E_{\mu_0} \sup_{\theta \in N_{\theta^*}} \left\| \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g_i(\theta)}{\partial \theta'} \right) \right\| < \infty$; $E_{\mu_0} \sup_{\beta \in N_{\beta^*}} \left\| \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h_i(\beta)}{\partial \beta'} \right) \right\| < \infty$.

(e) *The matrix* $\left(I_{p_g} \otimes W_g E_{\mu_0} g_i(\theta^*)\right)' E_{\mu_0} \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g_i(\theta^*)}{\partial \theta'} \right)$ *is nonsingular; the matrix* $\left(I_{p_h} \otimes W_g E_{\mu_0} h_i(\beta^*)\right)' E_{\mu_0} \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h_i(\beta^*)}{\partial \beta'} \right)$ *is nonsingular.*

Assumption 11(a) implies that the weighted difference of $g_i(\theta^*)$ and $h_i(\beta^*)$ is not a degenerate random variable, i.e.

$$\Pr\left\{ g_i(\theta^*)' W_g E_{\mu_0} g_i(\theta^*) - h_i(\beta^*)' W_h E_{\mu_0} h_i(\beta^*) = 0 : \mu_0 \right\} = 0.$$

Assumption 11(b)-(d) assume that the moment functions g and h are sufficiently smooth in some neighborhoods of θ^* and β^* respectively, and the distribution of the data has sufficiently thin tails; they are similar to Assumption 2(f) of Kitamura (2000). Assumption 11(e) is a condition on the weighted matrices of the second derivatives of g and h.

We have the following result.

Theorem 12 (Approximate optimal test for nonlinear models) *Suppose that Assumptions 9 and 11 hold. Assume further that* $\left\| E_{\hat{p}^{(s-1)}} g_i\left(\hat{\theta}^{(s-1)}\right) - E_{\mu_0} g_i(\theta^*) \right\| = O_p\left(n^{-1/2}\right)$ *and* $\left\| E_{\hat{p}^{(s-1)}} h_i\left(\hat{\beta}^{(s-1)}\right) - E_{\mu_0} h_i(\beta^*) \right\| = O_p\left(n^{-1/2}\right)$ *for some* $s \in \mathbb{N}$. *Then,*

(a) $2nT_n^{G(s)} \to_d \chi_1^2$ *under* H_0, *and* $2nT_n^{G(s)} \to \infty$ *almost surely under* H_1,

(b) $\left\| E_{\hat{p}^{(s)}} g_i\left(\hat{\theta}^{(s)}\right) - E_{\mu_0} g_i\left(\theta^*\right) \right\| = O_p\left(n^{-1/2}\right)$ *and* $\left\| E_{\hat{p}^{(s)}} h_i\left(\hat{\beta}^{(s)}\right) - E_{\mu_0} h_i\left(\beta^*\right) \right\| = O_p\left(n^{-1/2}\right)$, *provided that* H_0 *is true.*

Remarks. **(a)** According to part (a) of the theorem, if $E_{\hat{p}^{(s-1)}} g_i\left(\hat{\theta}^{(s-1)}\right)$ and $E_{\hat{p}^{(s-1)}} h_i\left(\hat{\beta}^{(s-1)}\right)$ are $n^{-1/2}$ distant from $E_{\mu_0} g_i\left(\theta^*\right)$ and $E_{\mu_0} h_i\left(\beta^*\right)$, then the statistic computed after s iteration has a χ_1^2 asymptotic distribution under the null. Furthermore, at iteration s, the estimated expectations of g_i and h_i are also $n^{-1/2}$ distant from their true expected values, provided that H_0 is true.

(b) For $2nT_n^{G(s)}$ to have a χ_1^2 asymptotic null distribution at each iteration s, it is sufficient to initialize the algorithm with the starting values that are $n^{-1/2}$ distant from their corresponding population counterparts. Thus, one should pick starting values $\left(\hat{\theta}^{(0)}, \hat{\beta}^{(0)}, \hat{\lambda}^{(0)}, \hat{p}^{(0)}\right)$ satisfying $\left\| \sum_{i=1}^n \hat{p}_i^{(0)} g_i\left(\hat{\theta}^{(0)}\right) - E_{\mu_0} g_i\left(\theta^*\right) \right\| = O_p\left(n^{-1/2}\right)$, $\left\| \sum_{i=1}^n \hat{p}_i^{(0)} h_i\left(\hat{\beta}^{(0)}\right) - E_{\mu_0} h_i\left(\beta^*\right) \right\| = O_p\left(n^{-1/2}\right)$, and $\hat{\lambda}^{(0)} = o_p(1)$. For example, under mild regularity conditions, these conditions are satisfied by

$$\begin{aligned}
\hat{p}_i^{(0)} &= \frac{1}{n}, \quad \hat{\lambda}^{(0)} = 0, \\
\hat{\theta}^{(0)} &= \theta\left(\hat{p}^{(0)}\right) = \arg\inf_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n g_i(\theta) \right\|_{W_g}^2, \\
\hat{\beta}^{(0)} &= \beta\left(\hat{p}^{(0)}\right) = \arg\inf_{\beta \in B} \left\| \frac{1}{n} \sum_{i=1}^n h_i(\beta) \right\|_{W_h}^2.
\end{aligned}$$

(c) As we mention in the introduction, our result differs from that of WDB96. In their paper, WDB96 provide high level assumptions under which the iterative and ELR statistics are asymptotically equivalent; however, the distribution of the ELR statistic in the case of a nonlinear in probabilities constraint remains unknown. On the other hand, we derive the asymptotic null distribution of the iterated statistic at each iteration by relying on the more primitive assumptions. Furthermore, our result implies that the ELR statistic corresponding to the nonlinear in probabilities problem has a χ^2 asymptotic null distribution, because the sequence $T_n^{G(s)}$ converges to T_n^G when it converges.

6 Conclusion

In this paper, we consider optimality in model comparison hypothesis testing for misspecified unconditional moment restriction models. Based on the generalized Neyman-Pearson optimality criterion, which focuses on the convergence rates of the type I and II error probabilities under fixed global alternatives, we find an optimal test statistic that is defined by the Kullback-Leibler information criterion. We propose approximate test statistics to the optimal test for linear and nonlinear models. For linear instrumental variable regression models, we obtain the conventional empirical likelihood ratio test. For general nonlinear moment restrictions, we develop a new test statistic based on an iterative algorithm. The asymptotic properties are derived for these test statistics.

A Proofs

A.1 Proof of Theorem 5

First, we check Definition 4 (a). Without loss of generality, we set as $c = 2$ in (7). Pick any $\delta > 0$ and $P_0 \in \mathcal{P}_0$. We start by showing that for each $\delta' \in (0, \delta/2)$,

$$cl\left(\Lambda_{1,\delta}^{\delta}\right) \subset \Lambda_{1,\delta'}^{\delta'}. \tag{18}$$

Pick any and $\nu \in cl\left(\Lambda_{1,\delta}^{\delta}\right)$. It is sufficient for (18) to show that

$$\inf_{\mu \in \mathcal{P}_0} I\left(\nu' \| \mu\right) > \alpha \quad \text{for each } \nu' \in cl\left(B\left(\nu, 2\delta'\right)\right). \tag{19}$$

Since $\nu \in cl\left(\Lambda_{1,\delta}^{\delta}\right)$, there exists $\omega \in \mathcal{M}$ such that $D_L\left(\nu, \omega\right) \leq \delta + \left(\delta - 2\delta'\right)/2$ and $\inf_{\mu \in \mathcal{P}_0} I\left(\omega' \| \mu\right) > \alpha$ for each $\omega' \in cl\left(B\left(\omega, 2\delta\right)\right)$. Thus, it is sufficient for (19) to show that $\nu' \in cl\left(B\left(\omega, 2\delta\right)\right)$ for each $\nu' \in cl\left(B\left(\nu, 2\delta'\right)\right)$. This can be shown by the triangle inequality:

$$\begin{aligned} D_L\left(\nu', \omega\right) &\leq D_L\left(\nu', \nu\right) + D_L\left(\nu, \omega\right) \\ &\leq 2\delta' + \delta + \left(\delta - 2\delta'\right)/2 \\ &< 2\delta, \end{aligned}$$

for each $\nu' \in cl\left(B\left(\nu, 2\delta'\right)\right)$. Therefore, we obtain (18). Now, observe that

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log \Pr\left\{\mu_n \in \Lambda_{1,\delta}^{\delta} : P_0\right\} \\
&\leq \limsup_{n\to\infty}\frac{1}{n}\log \Pr\left\{\mu_n \in cl\left(\Lambda_{1,\delta}^{\delta}\right) : P_0\right\} \\
&\leq -\inf_{P\in cl\left(\Lambda_{1,\delta}^{\delta}\right)} I\left(P\|P_0\right) \\
&\leq -\inf_{P\in \Lambda_{1,\delta'}^{\delta'}} I\left(P\|P_0\right) \\
&\leq -\alpha,
\end{aligned}$$

where the first inequality follows from a set inclusion relationship, the second inequality follows from Sanov's theorem, the third inequality follows from (18), and the last inequality follows from the definition of $\Lambda_{1,\delta'}^{\delta'}$. Therefore, the test Λ satisfies Definition 4 (a).

We now check Definition 4 (b). Without loss of generality, we set as $\bar{\delta} = 6\delta$. Pick any $\delta > 0$ and $P_1 \in \mathcal{M} \setminus \mathcal{P}_0$. We start by showing that

$$\Lambda_{0,2.1\delta}^{2.1\delta} \subset \Omega_0^{\delta}. \tag{20}$$

Suppose otherwise. Then there exists a sequence $\{\xi_m\}_{m\in\mathbb{N}}$ such that $\xi_m \in \Lambda_{0,2.1\delta}^{2.1\delta}$ and $\xi_m \in \Omega_1^{\delta}$ for all $m \in \mathbb{N}$. Since $\xi_m \in \Lambda_{0,2.1\delta}^{2.1\delta}$, there exists $\{\xi'_m\}_{m\in\mathbb{N}}$ such that $D_L\left(\xi_m, \xi'_m\right) < 4.2\delta$ and $\inf_{\mu\in\mathcal{P}_0} I\left(\xi'_m\|\mu\right) \leq \alpha$. The set $\{\xi \in \mathcal{M} : \inf_{\mu\in\mathcal{P}_0} I\left(\xi\|\mu\right) \leq \alpha\}$ is assumed to be compact, and therefore there exists a subsequence $\left\{\xi'_{m_k}\right\}_{k\in\mathbb{N}}$ such that $\xi'_{m_k} \to \xi' \in \{\xi \in \mathcal{M} : \inf_{\mu\in\mathcal{P}_0} I\left(\xi\|\mu\right) \leq \alpha\}$ as $k \to \infty$. Also, from $\xi_m \in \Omega_1^{\delta}$ and $D_L\left(\xi_m, \xi'_m\right) < 4.2\delta$ for all $m \in \mathbb{N}$, we have $\xi'_{m_k} \in \Omega_1^{5.2\delta}$ and thus $B\left(\xi'_{m_k}, \delta/2\right) \subset \Omega_1^{6\delta}$ for all $k \in \mathbb{N}$, which implies that the limit ξ' satisfies $B\left(\xi', \delta/4\right) \subset \Omega_1^{6\delta}$. Thus, Sanov's theorem implies

$$\begin{aligned}
&\sup_{P_0\in\mathcal{P}_0}\liminf_{n\to\infty}\frac{1}{n}\log \Pr\left\{\mu_n \in \Omega_1^{6\delta} : P_0\right\} \\
&\geq \sup_{P_0\in\mathcal{P}_0}\liminf_{n\to\infty}\frac{1}{n}\log \Pr\left\{\mu_n \in B\left(\xi', \delta/4\right) : P_0\right\} \\
&\geq -\inf_{P_0\in\mathcal{P}_0}\inf_{P\in B(\xi',\delta/4)} I\left(P\|P_0\right) \\
&\geq -\alpha.
\end{aligned}$$

Since this contradicts with the requirement for Ω, we obtain (20). Now, observe that

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in\Omega_0^\delta : P_1\right\}\\
&\geq \limsup_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in\Lambda_{0,2.1\delta}^{2.1\delta} : P_1\right\}\\
&\geq \liminf_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in int\left(\Lambda_{0,2.1\delta}^{2.1\delta}\right) : P_1\right\}\\
&\geq -\inf_{P\in int\left(\Lambda_{0,2.1\delta}^{2.1\delta}\right)} I\left(P\|P_1\right)\\
&\geq -\inf_{P\in\Lambda_{0,\delta}^{\delta}} I\left(P\|P_1\right)\\
&\geq \limsup_{n\to\infty}\frac{1}{n}\log\Pr\left\{\mu_n\in\Lambda_{0,\delta}^{\delta} : P_1\right\},
\end{aligned}$$

where the first inequality follows from (20), the second inequality follows from a set inclusion relationship, the third inequality follows from Sanov's theorem, the fourth inequality follows from (18), and the last inequality follows from Sanov's theorem. Therefore, the test Λ satisfies Definition 4 (b). ■

A.2 Proof of Lemma 10

Consider the optimization problem in (14); its Lagrangian is

$$\begin{aligned}
\mathcal{L} = &\sum_{i=1}^{n}\log\left(np_i\right)+\gamma\left(\sum_{i=1}^{n}p_i-1\right)\\
&-\frac{n\lambda}{2}\left(\sum_{i=1}^{n}p_i g_i\left(\theta\left(p\right)\right)\right)' W_g\left(\sum_{i=1}^{n}p_i g_i\left(\theta\left(p\right)\right)\right)\\
&+\frac{n\lambda}{2}\left(\sum_{i=1}^{n}p_i h_i\left(\beta\left(p\right)\right)\right)' W_h\left(\sum_{i=1}^{n}p_i h_i\left(\beta\left(p\right)\right)\right).
\end{aligned}$$

The first-order condition corresponding to p_i is

$$\begin{aligned}
0 = \hat{p}_i^{-1}+\hat{\gamma}-n\hat{\lambda}\Bigg[&\left(g_i\left(\theta\left(\hat{p}\right)\right)+\sum_{j=1}^{n}\hat{p}_j\frac{\partial g_j\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\frac{\partial\theta\left(\hat{p}\right)}{\partial p_i}\right)' W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)\\
&-\left(h_i\left(\beta\left(\hat{p}\right)\right)+\sum_{j=1}^{n}\hat{p}_j\frac{\partial h_j\left(\beta\left(\hat{p}\right)\right)}{\partial\beta'}\frac{\partial\beta\left(\hat{p}\right)}{\partial p_i}\right)' W_h\left(\sum_{j=1}^{n}\hat{p}_j h_j\left(\beta\left(\hat{p}\right)\right)\right)\Bigg]. \quad (21)
\end{aligned}$$

The first-order conditions for $\theta(\hat{p})$ and $\beta(\hat{p})$ are

$$\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial g_j\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\right)'W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)=0, \tag{22}$$

$$\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial h_j\left(\beta\left(\hat{p}\right)\right)}{\partial\beta'}\right)'W_h\left(\sum_{j=1}^{n}\hat{p}_j h_j\left(\beta\left(\hat{p}\right)\right)\right)=0. \tag{23}$$

Thus, by the implicit function theorem, the derivatives $\partial\theta(p)/\partial p_i$ and $\partial\beta(p)/\partial p_i$ can be obtained as

$$\begin{aligned}\frac{\partial\theta\left(\hat{p}\right)}{\partial p_i}=&-H_g^{-1}\left(\hat{p},\theta\left(\hat{p}\right)\right)\left(\frac{\partial g_i\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\right)'W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)-\\&-H_g^{-1}\left(\hat{p},\theta\left(\hat{p}\right)\right)\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial g_j\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\right)'W_g g_i\left(\theta\left(\hat{p}\right)\right).\end{aligned}$$

From Assumption 9(d), $\partial\theta(\hat{p})/\partial p_i$ exists with probability approaching one. We can obtain a similar expression for $\partial\beta(\hat{p})/\partial p_i$.

Next, note that by (22),

$$\begin{aligned}&\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial g_j\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\frac{\partial\theta\left(\hat{p}\right)}{\partial p_i}\right)'W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)\\&=\frac{\partial\theta\left(\hat{p}\right)'}{\partial p_i}\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial g_j\left(\theta\left(\hat{p}\right)\right)}{\partial\theta'}\right)'W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)\\&=0,\end{aligned}$$

and similarly,

$$\left(\sum_{j=1}^{n}\hat{p}_j\frac{\partial h_j\left(\beta\left(\hat{p}\right)\right)}{\partial\beta'}\frac{\partial\beta\left(\hat{p}\right)}{\partial p_i}\right)'W_h\left(\sum_{j=1}^{n}\hat{p}_j h_j\left(\beta\left(\hat{p}\right)\right)\right)=0.$$

Thus, the first-order condition for $\hat{p}$ in (21) simplifies to

$$\hat{p}_i^{-1}=\hat{\gamma}-n\hat{\lambda}\left[g_i\left(\theta\left(\hat{p}\right)\right)'W_g\left(\sum_{j=1}^{n}\hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right)-h_i\left(\beta\left(\hat{p}\right)\right)'W_h\left(\sum_{j=1}^{n}\hat{p}_j h_j\left(\beta\left(\hat{p}\right)\right)\right)\right].$$

By multiplying $\hat{p}_i^{-1}$ and taking summation to the both sides, we have $\hat{\gamma} = -n$ and then

$$n^{-1}\hat{p}_i^{-1} = 1 + \hat{\lambda} d_i \left(\theta\left(\hat{p}\right), \beta\left(\hat{p}\right); \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right), \tag{24}$$

where $\hat{\lambda}$ solves

$$\sum_{i=1}^{n} \frac{d_i \left(\theta\left(\hat{p}\right), \beta\left(\hat{p}\right); \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right)}{1 + \hat{\lambda} d_i \left(\theta\left(\hat{p}\right), \beta\left(\hat{p}\right); \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right)} = 0. \tag{25}$$

Suppose $\hat{\theta}_d$, $\hat{\beta}_d$, and $\hat{\lambda}_d$ solve the minimax problem in (15). Further, define $\hat{p}_{d,i}$ such that

$$n^{-1}\hat{p}_{d,i}^{-1} = 1 + \hat{\lambda}_d d_i \left(\hat{\theta}_d, \hat{\beta}_d; \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right).$$

The first-order conditions for $\hat{\theta}_d$, $\hat{\beta}_d$, and $\hat{\lambda}_d$ are

$$\left(\sum_{j=1}^{n} \hat{p}_{d,j} \frac{\partial g_j\left(\hat{\theta}_d\right)}{\partial \theta'}\right)' W_g \left(\sum_{j=1}^{n} \hat{p}_j g_j\left(\theta\left(\hat{p}\right)\right)\right) = 0, \tag{26}$$

$$\left(\sum_{j=1}^{n} \hat{p}_{d,j} \frac{\partial h_j\left(\hat{\beta}_d\right)}{\partial \beta'}\right)' W_h \left(\sum_{j=1}^{n} \hat{p}_j h_j\left(\beta\left(\hat{p}\right)\right)\right) = 0, \tag{27}$$

$$\sum_{i=1}^{n} \frac{d_i \left(\hat{\theta}_d, \hat{\beta}_d; \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right)}{1 + \hat{\lambda}_d d_i \left(\hat{\theta}_d, \hat{\beta}_d; \hat{p}, \theta\left(\hat{p}\right), \beta\left(\hat{p}\right)\right)} = 0. \tag{28}$$

By comparing (22), (23), and (25) with (26)-(28), it follows that $\hat{\theta}$, $\hat{\beta}$, and $\hat{\lambda}$ solve the dual problem. Furthermore, by Assumption 9(b), the solution is unique with probability approaching one. ■

A.3 Proof of Theorem 12

Proof of (a). Pick any $s \in \mathbb{N}$. The following steps yield the conclusion:

$$\begin{aligned} 2nT_n^{G(s)} &= 2\sum_{i=1}^{n} \log\left(1 + \hat{\lambda}^{(s)} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right) \\ &= \frac{\left(\sum_{i=1}^{n} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right)^2}{\sum_{i=1}^{n} d_i^2 \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)} + o_p(1) \end{aligned} \tag{29}$$

$$= \frac{\left(\sum_{i=1}^{n} d_i\left(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*\right)\right)^2}{E_{\mu_0} d_i^2\left(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*\right)} + o_p(1) \tag{30}$$

$$\to_d \chi_1^2. \tag{31}$$

First, we show (29). An expansion of the first-order condition for $\hat{\lambda}^{(s)}$ around $\hat{\lambda}^{(s)} = 0$ yields

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \frac{d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)}{1 + \hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)} \\ &= \sum_{i=1}^{n} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \\ &\quad - \hat{\lambda}^{(s)} \sum_{i=1}^{n} \frac{d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)^2}{\left(1 + \bar{\lambda} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right)^2}, \end{aligned} \tag{32}$$

where $\bar{\lambda}$ is a point on the line joining $\hat{\lambda}^{(s)}$ and 0. Assumption 11(b) implies that

$$\max_{1 \le i \le n} \sup_{\theta \in N_{\theta^*}} \|g_i(\theta)\| = o_p\left(n^{1/2}\right), \tag{33}$$

$$\max_{1 \le i \le n} \sup_{\beta \in N_{\beta^*}} \|h_i(\beta)\| = o_p\left(n^{1/2}\right), \tag{34}$$

(see the proof of (2.4) on page 701 of Guggenberger and Smith (2005)). By Lemma 13, $\hat{\theta}^{(s)}$ and $\hat{\beta}^{(s)}$ are in the $n^{-1/2}$ neighborhoods of θ^* and β^* respectively, and $\hat{\lambda}^{(s)} = O_p\left(n^{-1/2}\right)$. It follows now by (33) and (34),

$$\left|\bar{\lambda}\right| \max_{1 \le i \le n} \left| d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right| = o_p(1). \tag{35}$$

Also note that by a standard consistency argument using the assumptions of the theorem on $E_{\hat{p}^{(s-1)}} g_i\left(\hat{\theta}^{(s-1)}\right)$ and $E_{\hat{p}^{(s-1)}} h_i\left(\hat{\beta}^{(s-1)}\right)$, Lemma 13(a), Assumption 11(b)-(c), and Cauchy-Schwarz inequality:

$$\frac{1}{n} \sum_{i=1}^{n} d_i^2\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \xrightarrow{p} E_{\mu_0} d_i^2\left(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*\right), \tag{36}$$

Thus, under (35) and (36) with Assumption 11(a), we can solve (32) for $\hat{\lambda}^{(s)}$ as

$$\hat{\lambda}^{(s)} = \frac{\sum_{i=1}^{n} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)}{\sum_{i=1}^{n} d_i^2 \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)} + o_p(1), \tag{37}$$

with probability approaching one. On the other hand, an expansion of $2\sum_{i=1}^{n} \log\left(1 + \hat{\lambda}^{(s)} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right)$ around $\hat{\lambda}^{(s)} = 0$ yields

$$\begin{aligned}
&2\sum_{i=1}^{n} \log\left(1 + \hat{\lambda}^{(s)} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\right) \\
&= 2\hat{\lambda}^{(s)} \sum_{i=1}^{n} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \\
&- \hat{\lambda}^{(s)2} \sum_{i=1}^{n} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)^2 \\
&+ \frac{2}{3}\hat{\lambda}^{(s)3} \sum_{i=1}^{n} \frac{d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)^3}{1 + \dot{\lambda} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)},
\end{aligned} \tag{38}$$

where $\dot{\lambda}$ is a point on the line joining $\hat{\lambda}^{(s)}$ and 0. Ignoring the constant, the absolute value of the third term in (38) is bounded by

$$\begin{aligned}
&\left|\hat{\lambda}^{(s)}\right|^3 \max_{1\le i\le n} \left| \frac{1}{1 + \dot{\lambda} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)} \right| \\
&\times \max_{1\le i\le n} \left| d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \right| \\
&\times \left| \sum_{i=1}^{n} d_i \left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)^2 \right| \\
&= O_p\left(n^{-3/2}\right) O_p(1)\, o_p\left(n^{1/2}\right) O_p(n) = o_p(1),
\end{aligned}$$

where the equality follows from Lemma 13(b), (35), and (36). Therefore, from (37) and (38), we obtain (29).

Next, we show (30). It is sufficient to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\mu_0,\theta^*,\beta^*\right)\xrightarrow{p}0. \quad (39)$$

Observe that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\mu_0,\theta^*,\beta^*\right)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\theta^*,\beta^*\right)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},,\theta^*,\beta^*\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\mu_0,\theta^*,\beta^*\right)\\
&=T_1+T_2+T_3.
\end{aligned}$$

For T_1, an expansion around $\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)}\right)=(\theta^*,\beta^*)$ yields

$$\begin{aligned}
T_1&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)}{\partial\theta'}\left(\hat{\theta}^{(s)}-\theta^*\right)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)}{\partial\beta'}\left(\hat{\beta}^{(s)}-\beta^*\right)+o_p(1) \quad (40)\\
&=\left(\sum_{j=1}^{n}\hat{p}_j^{(s-1)}g_j\left(\hat{\theta}^{(s-1)}\right)\right)'W_g\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\theta^*)}{\partial\theta'}\right)\sqrt{n}\left(\hat{\theta}^{(s)}-\theta^*\right)\\
&+\left(\sum_{j=1}^{n}\hat{p}_j^{(s-1)}h_j\left(\hat{\beta}^{(s-1)}\right)\right)'W_h\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial h_i(\beta^*)}{\partial\beta'}\right)\sqrt{n}\left(\hat{\beta}^{(s)}-\beta^*\right)+o_p(1)\\
&=\left(E_{\mu_0}g_i(\theta^*)\right)'W_g\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\theta^*)}{\partial\theta'}\right)\sqrt{n}\left(\hat{\theta}^{(s)}-\theta^*\right)\\
&+\left(E_{\mu_0}h_i(\beta^*)\right)'W_h\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial h_i(\beta^*)}{\partial\beta'}\right)\sqrt{n}\left(\hat{\beta}^{(s)}-\beta^*\right)+o_p(1)\\
&=\left(E_{\mu_0}g_i(\theta^*)\right)'W_g\left(E_{\mu_0}\frac{\partial g_i(\theta^*)}{\partial\theta'}\right)\sqrt{n}\left(\hat{\theta}^{(s)}-\theta^*\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \left(E_{\mu_0} h_i\left(\beta^*\right)\right)' W_h \left(E_{\mu_0} \frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\right) \sqrt{n}\left(\hat{\beta}^{(s)} - \beta^*\right) + o_p(1) \\
&= o_p(1),
\end{aligned}
$$

where the second equality follows from the definition of $d_i\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)$, the third equality follows from the conditions of the theorem on $E_{\hat{p}^{(s-1)}} g_i\left(\hat{\theta}^{(s-1)}\right)$ and $E_{\hat{p}^{(s-1)}} h_i\left(\hat{\beta}^{(s-1)}\right)$, the fourth equality follows from the weak law of large numbers, and the last equality follows from the first order conditions of θ^* and β^*. The reminder term in (40) is bounded by

$$
\begin{aligned}
&\|W_g\| \left\| E_{\hat{p}^{(s-1)}} g_i\left(\hat{\theta}^{(s-1)}\right)\right\| \left\|\hat{\theta}^{(s)} - \theta^*\right\|^2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\theta \in N_{\theta^*}} \left\| \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g_i(\theta)}{\partial \theta'}\right)\right\| \\
&+ \|W_h\| E_{\hat{p}^{(s-1)}} h_i\left(\hat{\beta}^{(s-1)}\right) \left\|\hat{\beta}^{(s)} - \beta^*\right\|^2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sup_{\beta \in N_{\beta^*}} \left\| \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h_i(\beta)}{\partial \beta'}\right)\right\|,
\end{aligned}
$$

which is $o_p(1)$ by Lemma 13(a) and Assumption 11(d).

Similarly, for T_2, an expansion around $\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) = (\theta^*, \beta^*)$ yields

$$
\begin{aligned}
T_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial d_i\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \theta^*, \beta^*\right)}{\partial \theta'} \left(\hat{\theta}^{(s-1)} - \theta^*\right) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial d_i\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \theta^*, \beta^*\right)}{\partial \beta'} \left(\hat{\beta}^{(s-1)} - \beta^*\right) + o_p(1) \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} g_i\left(\theta^*\right)\right)' W_g \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} \frac{\partial g_j\left(\theta^*\right)}{\partial \theta'}\right) \sqrt{n}\left(\hat{\theta}^{(s-1)} - \theta^*\right) \\
&+ \left(\frac{1}{n} \sum_{i=1}^{n} h_i\left(\beta^*\right)\right)' W_h \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} \frac{\partial h_j\left(\beta^*\right)}{\partial \beta'}\right) \sqrt{n}\left(\hat{\beta}^{(s)} - \beta^*\right) + o_p(1) \\
&= \left(E_{\mu_0} g_i\left(\theta^*\right)\right)' W_g \left(E_{\mu_0} \frac{\partial g_i\left(\theta^*\right)}{\partial \theta'}\right) \sqrt{n}\left(\hat{\theta}^{(s-1)} - \theta^*\right) \\
&+ \left(E_{\mu_0} h_i\left(\beta^*\right)\right)' W_h \left(E_{\mu_0} \frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\right) \sqrt{n}\left(\hat{\beta}^{(s)} - \beta^*\right) + o_p(1) \\
&= o_p(1),
\end{aligned}
$$

where the second equality follows from the definition of $d_i\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \theta^*, \beta^*\right)$, and the last equality follows from the first order conditions of θ^* and β^*. For the third

equality, using the definition of $\hat{p}^{(s)}$,

$$\begin{aligned}
&\left\| \sum_{j=1}^{n} \hat{p}_j^{(s-1)} \frac{\partial g_j(\theta^*)}{\partial \theta'} - \frac{1}{n} \sum_{j=1}^{n} \frac{\partial g_j(\theta^*)}{\partial \theta'} \right\| \\
&\leq \left| \hat{\lambda}^{(s-1)} \right| \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{d_j\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}; \hat{p}^{(s-2)}, \hat{\theta}^{(s-2)}, \hat{\beta}^{(s-2)}\right)}{1 + \hat{\lambda}^{(s-1)} d_j\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}; \hat{p}^{(s-2)}, \hat{\theta}^{(s-2)}, \hat{\beta}^{(s-2)}\right)} \frac{\partial g_i(\theta^*)}{\partial \theta'} \right\| \\
&= \left| \hat{\lambda}^{(s-1)} \right| \sup_{1 \leq i \leq n} \left| 1 + \hat{\lambda}^{(s-1)} d_i\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s)}; \hat{p}^{(s-2)}, \hat{\theta}^{(s-2)}, \hat{\beta}^{(s-2)}\right) \right|^{-1} \\
&\times \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\partial g_i(\theta^*)}{\partial \theta'} \right\|^2 \frac{1}{n} \sum_{i=1}^{n} d_i^2\left(\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}; \hat{p}^{(s-2)}, \hat{\theta}^{(s-2)}, \hat{\beta}^{(s-2)}\right). \qquad (41)
\end{aligned}$$

Similarly to (35), we have

$$\left| \hat{\lambda}^{(s)} \right| \sup_{1 \leq i \leq n} \left| d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \right| = o_p(1),$$

and thus, by Assumption 11(c) and (36),

$$\left\| \sum_{j=1}^{n} \hat{p}_j^{(s-1)} \frac{\partial g_j(\theta^*)}{\partial \theta'} - \frac{1}{n} \sum_{j=1}^{n} \frac{\partial g_j(\theta^*)}{\partial \theta'} \right\| \leq O_p\left(n^{-1/2}\right) O_p(1) O_p(1) O_p(1).$$

By this and the weak law of large numbers,

$$\sum_{j=1}^{n} \hat{p}_j^{(s-1)} \frac{\partial g_j(\theta^*)}{\partial \theta'} = E_{\mu_0} \frac{\partial g_j(\theta^*)}{\partial \theta'} + o_p(1),$$

and a similar result holds for the h model.

For T_3, as above, we have that $\left\| \sum_{i=1}^{n} \hat{p}_i^{(s)} g_i(\theta^*) - n^{-1} \sum_{i=1}^{n} g_i(\theta^*) \right\|$ is bounded by

$$\begin{aligned}
&\left| \hat{\lambda}^{(s)} \right| \sup_{1 \leq i \leq n} \left| 1 + \hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)}, \hat{\beta}^{(s)}; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \right|^{-1} \\
&\times n^{-1} \sum_{i=1}^{n} \left\| g_i(\theta^*) \right\|^2 \sup_{\theta \in N_{\theta^*}} \sup_{\beta \in N_{\beta^*}} n^{-1} \sum_{i=1}^{n} d_i^2\left(\theta, \beta; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right) \\
&= O_p\left(n^{-1/2}\right) O_p(1) O_p(1) O_p(1).
\end{aligned}$$

Thus,

$$\sum_{i=1}^{n} \hat{p}_i^{(s)} g_i(\theta^*) = E_{\mu_0} g_i(\theta^*) + O_p\left(n^{-1/2}\right), \tag{42}$$
$$\sum_{i=1}^{n} \hat{p}_i^{(s)} h_i(\beta^*) = E_{\mu_0} h_i(\beta^*) + O_p\left(n^{-1/2}\right),$$

and

$$\begin{aligned} T_3 = & \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\theta^*)\right)' W_g \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} g_j(\theta^*) - E_{\mu_0} g_i(\theta^*)\right) \\ & + \left(\frac{1}{n}\sum_{i=1}^{n} h_i(\beta^*)\right)' W_h \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} h_j(\beta^*) - E_{\mu_0} h_i(\beta^*)\right) \\ & = o_p(1). \end{aligned}$$

Therefore, we obtain (39).

Lastly, for (31), since according to H_0, $E_{\mu_0} d_i(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*) = 0$, by Assumption 11(a) and a central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*) \xrightarrow{d} N\left(0, E_{\mu_0} d_i^2(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*)\right).$$

On the other hand, under H_1, the weak law of large numbers yields

$$\frac{1}{n}\sum_{i=1}^{n} d_i(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*) \xrightarrow{p} E_{\mu_0} d_i(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*) \neq 0,$$

and this implies $2nT_n^{G(s)} \to \infty$ with probability one.

Proof of (b). At iteration s we have the following first order condition:

$$\begin{aligned} \sum_{i=1}^{n} \hat{p}_j^{(s)} \Bigg[& g_i\left(\hat{\theta}^{(s)}\right)' W_g \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} g_j\left(\hat{\theta}^{(s-1)}\right)\right) - \\ & - h_i\left(\hat{\beta}^{(s)}\right)' W_h \left(\sum_{j=1}^{n} \hat{p}_j^{(s-1)} h_j\left(\hat{\beta}^{(s-1)}\right)\right)\Bigg] = 0. \end{aligned} \tag{43}$$

From (43), the null restriction $\|E_{\mu_0} g_i(\theta^*)\|_{W_g} = \|E_{\mu_0} h_i(\beta^*)\|_{W_h}$, and by the assump-

tions of the theorem:

$$\left(\sum_{i=1}^{n}\hat{p}_i^{(s)}g_i\left(\hat{\theta}^{(s)}\right)-E_{\mu_0}g_1\left(\theta^*\right)\right)'W_gE_{\mu_0}g_1\left(\theta^*\right)-$$
$$-\left(\sum_{i=1}^{n}\hat{p}_i^{(s)}h_i\left(\hat{\beta}^{(s)}\right)-E_{\mu_0}h_1\left(\beta^*\right)\right)'W_hE_{\mu_0}h_1\left(\beta^*\right)=O_p(n^{-1/2}). \quad (44)$$

Using an expansion of $g_i\left(\hat{\theta}^{(s)}\right)$ around $g_i\left(\theta^*\right)$, we obtain

$$\left|\left(\sum_{i=1}^{n}\hat{p}_i^{(s)}g_i\left(\hat{\theta}^{(s)}\right)-E_{\mu_0}g_1\left(\theta^*\right)\right)'W_gE_{\mu_0}g_1\left(\theta^*\right)\right|$$
$$\leq\|W_g\|\,\|E_{\mu_0}g_1\left(\theta^*\right)\|\left\|\sum_{i=1}^{n}\hat{p}_i^{(s)}g_i\left(\theta^*\right)-E_{\mu_0}g_1\left(\theta^*\right)\right\|$$
$$+\|W_g\|\,\|E_{\mu_0}g_1\left(\theta^*\right)\|\left\|\hat{\theta}^{(s)}-\theta^*\right\|\sup_{\theta\in N_{\theta^*}}\sum_{j=1}^{n}\hat{p}_j^{(s)}\left\|\frac{\partial g_j\left(\theta\right)}{\partial\theta'}\right\|. \quad (45)$$

Next, using (42) for the first summand and an argument similar to that in (41) and Assumption 11(c) for the second summand, the right-hand side of (45) is $O_p\left(n^{-1/2}\right)$. We obtain that the first term on the left-hand side of (44) is $O_p\left(n^{-1/2}\right)$. It follows now that

$$\left(\sum_{i=1}^{n}\hat{p}_i^{(s)}h_i\left(\hat{\beta}^{(s)}\right)-E_{\mu_0}h_1\left(\beta^*\right)\right)'W_hE_{\mu_0}h_1\left(\beta^*\right)=O_p(n^{-1/2}),$$

or, since W_h has a full rank,

$$\left\|\sum_{i=1}^{n}\hat{p}_i^{(s)}h_i\left(\hat{\beta}^{(s)}\right)-E_{\mu_0}h_1\left(\beta^*\right)\right\|=O_p\left(n^{-1/2}\right).$$

By similar arguments and using (43), we can show that

$$\left\|\sum_{i=1}^{n}\hat{p}_i^{(s)}g_i\left(\hat{\theta}^{(s)}\right)-E_{\mu_0}g_1\left(\theta^*\right)\right\|=O_p\left(n^{-1/2}\right).$$

■

B Auxiliary Lemma

Lemma 13 *Under the assumptions of Theorem 12 and H_0, we have the following results.*

(a) $\left\|\hat{\theta}^{(s)}-\theta^*\right\| = O_p\left(n^{-1/2}\right)$ *and* $\left\|\hat{\beta}^{(s)}-\beta^*\right\| = O_p\left(n^{-1/2}\right)$.

(b) $\hat{\lambda}^{(s)} = O_p\left(n^{-1/2}\right)$.

Proof of (a). The first-order conditions at iteration s are:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\partial g_i\left(\hat{\theta}^{(s)}\right)'/\partial\theta\right) W_g E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right)}{1+\hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)}, \tag{46}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\partial h_i\left(\hat{\beta}^{(s)}\right)'/\partial\beta\right) W_h E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right)}{1+\hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)}, \tag{47}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{g_i\left(\hat{\theta}^{(s)}\right)' W_g E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right) - h_i\left(\hat{\beta}^{(s)}\right)' W_h E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right)}{1+\hat{\lambda}^{(s)} d_i\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)};\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right)}. \tag{48}$$

Let $\delta_n = \left\|\hat{\theta}^{(s)}-\theta^*\right\| + \left\|\hat{\beta}^{(s)}-\beta^*\right\| + \left\|\hat{\lambda}^{(s)}\right\|$. Expanding the first-order conditions in (46)-(48) around $\left(\hat{\theta}^{(s)},\hat{\beta}^{(s)},\hat{\lambda}^{(s)}\right) = (\theta^*,\beta^*,0)$, we obtain:

$$\begin{aligned}
o_p(\delta_n) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i(\theta^*)'}{\partial\theta} W_g E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right) + \\
&\quad + \left(I_{p_g} \otimes W_g E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right)\right)' \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta'} vec\left(\frac{\partial g_i(\theta^*)}{\partial\theta'}\right)\left(\hat{\theta}^{(s)}-\theta^*\right), \\
o_p(\delta_n) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i(\beta^*)'}{\partial\beta} W_h E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right) + \\
&\quad + \left(I_{p_h} \otimes W_h E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right)\right)' \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\beta'} vec\left(\frac{\partial h_i(\beta^*)}{\partial\beta'}\right)\left(\hat{\beta}^{(s)}-\beta^*\right), \\
o_p(\delta_n) &= \frac{1}{n}\sum_{i=1}^{n} d_i\left(\theta^*,\beta^*;\hat{p}^{(s-1)},\hat{\theta}^{(s-1)},\hat{\beta}^{(s-1)}\right) +
\end{aligned}$$

$$+ E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right)' W_g \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i\left(\theta^*\right)}{\partial \theta'}\left(\hat{\theta}^{(s)} - \theta^*\right)$$
$$- E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right)' W_h \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\left(\hat{\beta}^{(s)} - \beta^*\right)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} d_i^2\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\hat{\lambda}^{(s)}.$$

By Assumptions 9 (b) and 11 (c) and (d), and since by the assumptions of the lemma, $E_{\hat{p}^{(s-1)}} g_1\left(\hat{\theta}^{(s-1)}\right)$ and $E_{\hat{p}^{(s-1)}} h_1\left(\hat{\beta}^{(s-1)}\right)$ are $n^{-1/2}$ distant from their respective true values, the above equations yield

$$o_p\left(\delta_n\right) + O_p\left(n^{-1/2}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i\left(\theta^*\right)'}{\partial \theta} W_g E_{\mu_0} g_1\left(\theta^*\right) \tag{49}$$
$$+ \left(I_p \otimes W_g E_{\mu_0} g_1\left(\theta^*\right)\right)' \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g_i\left(\theta^*\right)}{\partial \theta'}\right)\left(\hat{\theta}^{(s)} - \theta^*\right),$$
$$o_p\left(\delta_n\right) + O_p\left(n^{-1/2}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i\left(\beta^*\right)'}{\partial \beta} W_h E_{\mu_0} h_1\left(\beta^*\right) \tag{50}$$
$$+ \left(I_r \otimes W_g E_{\mu_0} h_1\left(\beta^*\right)\right)' \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\right)\left(\hat{\beta}^{(s)} - \beta^*\right),$$
$$o_p\left(\delta_n\right) + O_p\left(n^{-1/2}\right) = \frac{1}{n}\sum_{i=1}^{n} d_i\left(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*\right) \tag{51}$$
$$+ E_{\mu_0} g_1\left(\theta^*\right)' W_g \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g_i\left(\theta^*\right)}{\partial \theta'}\left(\hat{\theta}^{(s)} - \theta^*\right)$$
$$- E_{\mu_0} h_1\left(\beta^*\right)' W_h \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\left(\hat{\beta}^{(s)} - \beta^*\right)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} d_i^2\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)\hat{\lambda}^{(s)}.$$

By the population first-order conditions for θ^* and β^* and Assumption 9(b),

$$E_{\mu_0} \frac{\partial g_i\left(\theta^*\right)'}{\partial \theta} W_g E_{\mu_0} g_i\left(\theta^*\right) = 0, \tag{52}$$
$$E_{\mu_0} \frac{\partial h_i\left(\beta^*\right)'}{\partial \beta} W_h E_{\mu_0} h_i\left(\beta^*\right) = 0.$$

Hence, the first terms on the right-hand sides of (49) and (50) are $O_p\left(n^{-1/2}\right)$. Further, by Assumption 11 (c)-(e), the matrices

$$\left(I_{p_g} \otimes W_g E_{\mu_0} g_1\left(\theta^*\right)\right)' \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta'} vec\left(\frac{\partial g_i\left(\theta^*\right)}{\partial \theta'}\right),$$

$$\left(I_{p_h} \otimes W_g E_{\mu_0} h_1\left(\beta^*\right)\right)' \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta'} vec\left(\frac{\partial h_i\left(\beta^*\right)}{\partial \beta'}\right),$$

are nonsingular and finite with probability approaching one. Thus, the conclusion follows from the same argument as on page 318 of Qin and Lawless (1994).

Proof of (b). Under H_0, $E_{\mu_0} d_i\left(\theta^*, \beta^*; \mu_0, \theta^*, \beta^*\right) = 0$, and therefore the first summand on the right-hand side of (51) is $O_p\left(n^{-1/2}\right)$. The second and third summands are $O_p\left(n^{-1}\right)$ by (52) and part (a) of this lemma. Further, by Assumption 11 (a) and (b),

$$\frac{1}{n} \sum_{i=1}^{n} d_i^2\left(\theta^*, \beta^*; \hat{p}^{(s-1)}, \hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}\right)$$

is strictly positive and finite with probability approaching one. The conclusion is followed by solving (51) for $\hat{\lambda}^{(s)}$. ■

References

Aguirregabiria, V., and P. Mira (2002): "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70(4), 1519–1543.

——— (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75(1), 1–53.

Chaganty, N. R., and R. L. Karandikar (1996): "Some Properties of the Kullback-Leibler Number," *Sankhyā Series A*, 58, 69–80.

Corradi, V., and N. R. Swanson (2007): "Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historic Data," *Journal of Econometrics*, 136(2), 699–723.

DAVIDSON, R., AND J. G. MACKINNON (1981): "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49(3), 781–793.

DEMBO, A., AND O. ZEITOUNI (1998): *Large Deviations Techniques and Applications*. Springer, New York.

DEUSCHEL, J. D., AND D. W. STROOCK (1989): *Large Deviations*. Academic Press, New York.

GUGGENBERGER, P., AND R. J. SMITH (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory*, 21(4), 667–709.

HALL, A. R., AND A. INOUE (2003): "The Large Sample Behavior of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114(2), 361–394.

HALL, P., AND B. LA SCALA (1990): "Methodology and Algorithms of Empirical Likelihood," *International Statistical Review*, 58(2), 109–127.

HANSEN, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

HOEFFDING, W. (1965): "Asymptotically Optimal Tests for Multinomial Distributions," *Annals of Mathematical Statistics*, 36(2), 369–408.

KASAHARA, H., AND K. SHIMOTSU (2008): "Pseudo-Likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models," *Journal of Econometrics*, forthcoming.

KITAMURA, Y. (2000): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," Working Paper, University of Pennsylvania.

——— (2001): "Asymptotic Optimality of Empirical Likelihood For Testing Moment Restrictions," *Econometrica*, 69(6), 1661–1672.

——— (2003): "A Likelihood-Based Approach to the Analysis of a Class of Nested and Non-Nested Models," Working Paper, University of Pennsylvania.

LEININGER, W. (1984): "A Generalization of the Maximum Theorem," *Economics Letters*, 15, 309–313.

MACKINNON, J. G. (1983): "Model Specification Tests Against Non-Nested Alternatives," *Econometric Reviews*, 2(1), 85–110.

NEWEY, W., AND R. J. SMITH (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

OTSU, T., AND Y.-J. WHANG (2008): "Testing for Non-nested Conditional Moment Restrictions via Conditional Empirical Likelihood," *Econometric Theory*, forthcoming.

OWEN, A. B. (2001): *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

PRESCOTT, E. C. (1991): "Real Business Cycle Theory: What Have We Learned?," *Revista de Análisis Económico*, 6(2), 3–19.

QIN, J., AND J. LAWLESS (1994): "Empirical likelihood and general estimating equations," *Annals of Statistics*, 22(1), 300–325.

RAMALHO, J. J. S., AND R. J. SMITH (2002): "Generalized Empirical Likelihood Non-Nested Tests," *Journal of Econometrics*, 107(1-2), 99–125.

RIVERS, D., AND Q. VUONG (2002): "Model Selection Tests For Nonlinear Dynamic Models," *Econometrics Journal*, 5(1), 1–39.

SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

SMITH, R. J. (1992): "Non-Nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 60(4), 973–980.

——— (1997): "Alternative Semi-parametric Likelihood Approaches to Generalised Method of Moments Estimation," *Economic Journal*, 107(441), 503–519.

VUONG, Q. H. (1989): "Likelihood Ratio Tests For Model Selection and Non-Nested Hypotheses," *Econometrica*, 57(2), 307–333.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25.

Wood, A. T. A., K. A. Do, and B. M. Broom (1996): "Sequential Linearization of Empirical Likelihood Constraints with Application to U-Statistics," *Journal of Computational and Graphical Statistics*, 5(4), 365–385.

Zeitouni, O., and M. Gutman (1991): "On Universal Hypotheses Testing via Large Deviations," *IEEE Transactions on Information Theory*, 37(2), 285–290.