A Nondegenerate Vuong Test*

Xiaoxia Shi†
University of Wisconsin - Madison
xiaoxia.shi@gmail.com
http://ssc.wisc.edu/~xshi

First Draft: April, 2009, Current Draft: February, 2014

Abstract

In this paper, I propose a one-step nondegenerate test as an alternative to the classical Vuong (1989) tests. I show that the new test achieves uniform asymptotic size control in both the overlapping and the nonoverlapping cases, while the classical Vuong tests do not. Meanwhile, the power of the new test can be substantially better than the two-step classical Vuong test and is not dominated by the one-step classical Vuong test. An extension to moment-based models is also developed. I apply the new test to the voter turnout data set of Coate and Conlin (2004) and find that it can yield model comparison conclusions different from those of the classical tests. The implementation of the new test is straightforward and can be done using the MATLAB and STATA routines accompanying this paper.

Keywords: Asymptotic size, Model comparison, Nonnested models, Voter Turnout, Vuong test

JEL classification number: C12, C52

1 Introduction

Vuong (1989) introduced tests of the hypothesis that two nonnested parametric models are equally distant in the Kullback-Leibler sense from the true data distribution. The tests have become important tools for comparing statistical models in empirical work. They have been generalized to moment-based models by Kitamura (2000), Rivers and Vuong (2002), Chen,

*A previous version was circulated under the title "The Asymptotic Size of the Vuong Test for Overlapping Models."

†I am deeply indebted to Donald Andrews for guidance and encouragement. I thank Bruce E. Hansen, Jack R. Porter, Quang H. Vuong, Frank Schorfheide and Kenneth D. West, Caterina Calsamiglia, Stephen Coate, Mike Conlin and three anonymous referees for many good comments and suggestions. The financial support from the University of Wisconsin-Madison via a Shoemaker Fellowship and via the Graduate School Fall Competition grants is gratefully acknowledged. Finally, I thank Wei Song for STATA coding. All errors are mine.

Hong, and Shum (2007), and Shi (2009), and to model comparisons based on mean-squared prediction error by Li (2009).

The classical Vuong framework mainly consists of two separate tests: a one-step test suggested for non-overlapping models and a two-step test suggested for overlapping models. In this paper, I find that both can generate severe size distortions in finite samples. I document the size distortions and then propose a new nondegenerate test that corrects the size distortions and applies universally to overlapping as well as nonoverlapping models. In addition, I show that the new test has good power and demonstrate using Monte Carlo simulations that its power can be substantially better than the two-step classical Vuong test and is comparable to the one-step classical Vuong test. The new test is straightforward to compute and its computation can be easily packaged into a MATLAB function or combined with the STATA maximum likelihood routine. Both the MATLAB and the STATA code files are provided with this paper.

The classical one-step Vuong test rejects the null hypothesis if the studentized log-likelihood ratio statistic (denoted by $\hat{T}_n$) exceeds a standard normal critical value. It has size distortion because the pointwise asymptotic distribution of $\hat{T}_n$ is discontinuous in a parameter ω_P^2 at $\omega_P^2 = 0$. Specifically, when $\omega_P^2 > 0$, the pointwise asymptotic distribution is $N(0,1)$, while when $\omega_P^2 = 0$, it is a nonstandard distribution (denoted by $J(0)$). In finite samples, the actual distribution of $\hat{T}_n$ is often close to a point *between* $N(0,1)$ and $J(0)$. If ω_P^2 is not sufficiently big, the point may be quite far from $N(0,1)$, causing the test to over-reject. The extreme case where ω_P^2 equals zero occurs when the best-fitting probability density functions (pdfs) of the candidate models are the same, which is possible when the models are overlapping. But the case where ω_P^2 is relatively small can occur for both overlapping and nonoverlapping models. In either case, the over-rejection may mislead the researcher into discarding a model that is not worse or even better than the model of comparison.[1]

The classical two-step Vuong test adds a pretest step where one tests $H_{00} : \omega_P^2 = 0$. The test rejects the null hypothesis only if both steps reject. The pretest can filter out some cases of small ω_P^2, but it does not filter out all cases. Thus, even though the pretest reduces size distortion, the reduction is often not enough. Moreover, the reduction may come at a large cost of power possibly due to the inefficiency in using a second moment to detect the difference between two pdfs, and to the negative correlation between $|\hat{T}_n|$ and the pretest statistic.[2] Monte Carlo Example 2 demonstrates the substantial power loss in the two-step Vuong test.

I formalize the discussion above using the local asymptotic theory, which is a widely used tool in the study of local power, weak instruments, etc. Specifically, I derive the asymptotic distribution of $\hat{T}_n$ under sequences of data-generating processes (DGPs) along which $n\omega_{P_n}^2 \to \sigma^2$

[1] It is important to note that when $n\omega_P^2$ is too small for $\hat{T}_n$ to be approximated by $N(0,1)$, the models may still yield rather different counterfactual predictions because the relatively small ω_P^2 may not be small in an absolute sense in the context of empirical work and because one of the candidate models may contain elements that are far from the other. Therefore, the stakes in drawing the correct model comparison conclusion can still be high.

[2] The pretest statistic is the denominator in $\hat{T}_n$.

for $\sigma^2 \in [0, \infty]$. The rate n is chosen so that the local asymptotic distribution obtained represents a smooth transition from $J(0)$ to $N(0,1)$ as σ^2 goes from 0 to ∞, which reflects our uncertainty about the size of ω_P^2. After obtaining the new asymptotic distribution, I examine it closely and find that a higher-order bias in the log-likelihood ratio statistic and a random term in its standard error are the key contributors to the size distortion of the classical Vuong tests. Based on this finding, I design a bias correction and a variance adjustment to $\hat{T}_n$. The new bias-corrected and variance-adjusted test statistic coupled with a simulation-based critical value forms my new test.

The classical Vuong tests were proposed for the comparison of parametric models based on the likelihood criterion. Yet the framework is flexible enough to be adapted to the comparison of other models based on other criteria. Kitamura (2000) extends the framework to moment equality models based on the relative entropy criterion. Rivers and Vuong (2002) compare moment equality models based on general criteria. Chen, Hong, and Shum (2007) compare a moment equality model with a parametric one. Li (2009) compares structural models based on simulated mean-squared error of prediction. Shi (2009) proposes tests for partially identified moment inequality models. In all of those papers except Li (2009), the discontinuity in the pointwise asymptotic distribution of the test statistics appears and the nonoverlapping case and the overlapping case are treated separately in the manner of Vuong (1989). Li (2009) does not have this discontinuity thanks to the simulated integrals used in his test statistic.

The analyses and the nondegenerate test proposed in this paper extend to the problem studied in Kitamura (2000) as well as other generalized empirical likelihood-based (GEL) model comparison problems. Section 6 is devoted to such an extension. I note that the nondegenerate test is even more appealing in practice in the semiparametric context. This is because for semiparametric models it is very difficult, if at all possible, to determine whether two models are overlapping or nonoverlapping. The nondegenerate test is robust to and can be applied in exactly the same way for both cases. Extensions to the other semiparametric settings mentioned above require derivations more substantially different and are left for future research. Also, this paper is not concerned with another important type of nonnested tests: the Cox-type tests (see, e.g., Cox (1961, 1962), Gourieroux and Monfort (1995), Otsu and Whang (2011) and Otsu, Seo, and Whang (2012)). Similar size distortions may occur with these tests and analyses similar to those of this paper are of potential interest.

A paper written concurrently with mine, Schennach and Wilhelm (2011), proposes a different test that also achieves uniform asymptotic size control. Their test is based on a sophisticated split-sample estimator of the log-likelihood ratio and the standard normal critical value. Because the critical value requires no simulation, their test is even easier to implement; in fact, it is as easy as the one-step Vuong test. However, my test has better power property because it does not rely on sample splitting. Monte Carlo simulations show that my test also achieves better finite sample size control.

The rest of the paper is organized as follows. Section 2 reviews the classical Vuong test and introduces the notation. Section 3 shows that the classical Vuong test over-rejects and discusses why. Section 4 proposes the new nondegenerate test and shows that it has correct null rejection rates uniformly over all DGPs. Section 5 shows that the new test has nontrivial power against $n^{-1/2}$-local alternatives and some n^{-1}-local alternatives. Section 6 extends the results in the previous sections to moment-based models. Section 7 demonstrates the finite sample performance of the classical, the nondegenerate, and the split-sample Vuong tests in a normal regression example and a joint normal location model example. Section 8 presents an empirical application to the comparison of voter turnout models. Section 9 concludes. Appendices A and B collect the proofs of all the formal results in the main text, and Supplemental Appendix C gives a third simulation example based on one-dimensional normal models.

2 Review of the Classical Vuong Tests

In this section, I review the classical Vuong tests and introduce the notation. Let $\{X_i \in \mathcal{X}\}_{i=1}^n = \{(Y_i', Z_i')' \in \mathcal{Y} \times \mathcal{Z}\}_{i=1}^n$ be observed i.i.d. data, where X_i is d_x-dimensional, Y_i and Z_i are d_y and d_z-dimensional, respectively, and $d_x = d_y + d_z$. Let the true distribution of X_i be P_0. Vuong (1989) considers the comparison between two parametric models, which are defined as parametric families of conditional densities of Y_i given Z_i:

$$\mathcal{F} = \{f_{Y|Z}(\cdot|\cdot;\theta) : \theta \in \Theta \subset R^{d_\theta}\} \text{ and } \mathcal{G} = \{g_{Y|Z}(\cdot|\cdot;\beta) : \beta \in \mathcal{B} \subset R^{d_\beta}\}. \tag{2.1}$$

Vuong (1989) tests the null hypothesis that the population log-likelihood ratio is zero:

$$H_0 : LR_{P_0} := E_{P_0}\Lambda_i(\phi_{P_0}^*) = 0, \tag{2.2}$$

where E_P denotes the expectation with respect to (w.r.t.) P, $\phi_P^* = (\theta_P^{*\prime}, \beta_P^{*\prime})'$ is the concatenated vector of the pseudo-true values: $\theta_P^* = \arg\max_{\theta\in\Theta} E_P \log f_{Y|Z}(Y_i|Z_i;\theta)$ and $\beta_P^* = \arg\max_{\beta\in\mathcal{B}} E_P \log g_{Y|Z}(Y_i|Z_i;\beta)$, and $\Lambda_i(\phi)$ is the logarithm of the ratio of the two pdfs:

$$\Lambda_i(\phi) = \log f_{Y|Z}(Y_i|Z_i;\theta) - \log g_{Y|Z}(Y_i|Z_i;\beta). \tag{2.3}$$

The population likelihood ratio LR_{P_0} is the difference in the Kullback-Leibler information distances from the true data distribution P_0 to the two models. Thus, H_0 can be interpreted as the models being equally distant from P_0 in the Kullback-Leibler sense. The null is tested against

$$H_1 : LR_{P_0} \neq 0, \tag{2.4}$$

which consists of two parts: $LR_{P_0} > 0$ and $LR_{P_0} < 0$. The former has the interpretation that P_0 is closer to model $\mathcal{F}$ than to $\mathcal{G}$ in the sense of Kullback-Leibler, while the latter has the opposite

interpretation.

Vuong (1989) distinguishes two types of testing scenarios based on the relationship between the candidate models. Specifically, models $\mathcal{F}$ and $\mathcal{G}$ are called *overlapping* if $\mathcal{F} \cap \mathcal{G} \neq \emptyset$, and *strictly nonnested* otherwise.[3] I review Vuong's (1989) treatment of the two scenarios in two subsections below. For notational simplicity, let $f_i(\theta) = f_{Y|Z}(Y_i|Z_i;\theta)$ and $g_i(\beta) = g_{Y|Z}(Y_i|Z_i;\beta)$.

2.1 One-Step Classical Vuong Test

When the models are strictly nonnested, Vuong (1989) proposes a one-step test that is based on the studentized log-likelihood ratio:

$$\begin{aligned} \hat{T}_n &= n^{1/2}\widehat{LR}_n/\hat{\omega}_n, \text{ where} \\ \widehat{LR}_n &= n^{-1}\sum_{i=1}^n \Lambda_i(\hat{\phi}_n) \text{ and } \hat{\omega}_n^2 = n^{-1}\sum_{i=1}^n \left[\Lambda_i(\hat{\phi}_n) - \widehat{LR}_n\right]^2, \end{aligned} \tag{2.5}$$

where $\hat{\phi}_n = (\hat{\theta}_n', \hat{\beta}_n')'$ and $\hat{\theta}_n$ and $\hat{\beta}_n$ are the maximum likelihood estimators from the two models respectively. Vuong (1989) shows that, under appropriate conditions and H_0,

$$\begin{aligned} n^{1/2}\widehat{LR}_n &\to_d N(0, \omega_{P_0}^2) \text{ and } \hat{\omega}_n^2 \to_p \omega_{P_0}^2, \text{ where} \\ \omega_P^2 &= \omega_P^2(\phi_P^*) \equiv E_P(\Lambda_i^2(\phi_P^*)) - (E_P\Lambda_i(\phi_P^*))^2. \end{aligned} \tag{2.6}$$

Based on these results, Vuong (1989) proposes to reject H_0 if and only if $|\hat{T}_n| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of $N(0,1)$. I refer to this test as the one-step classical Vuong test.

2.2 Two-Step Classical Vuong Test

When the models are overlapping, there is the possibility that $f_i(\theta_{P_0}^*) = g_i(\beta_{P_0}^*)$ almost surely because $\mathcal{F} \cap \mathcal{G} \neq \emptyset$. Recall that $\omega_{P_0}^2 = \log f_i(\theta_{P_0}^*) - \log g_i(\beta_{P_0}^*)$. Consequently, it is possible that $\omega_{P_0}^2 = 0$, in which case the results in (2.6) do not imply $\hat{T}_n \to_d N(0,1)$. The one-step test thus becomes invalid.

To guard against data distributions (P_0's) such that $\omega_{P_0}^2 = 0$, Vuong (1989) suggests a two-step test instead. In the first step, test

$$H_{00} : \omega_{P_0}^2 = 0. \tag{2.7}$$

If the first step does not reject, proceed no further and accept H_0. If the first step rejects, proceed to the one-step Vuong test. The first-step test (or the pretest) uses $n\hat{\omega}_n^2$ as the test statistic. Under H_{00}, Vuong (1989) shows that

$$n\hat{\omega}_n^2 \to_d Z'V^2Z, \tag{2.8}$$

[3]Note that by this definition, the "overlapping" case includes the "nested" case, where $\mathcal{F} \subseteq \mathcal{G}$ or $\mathcal{G} \subseteq \mathcal{F}$.

where $Z \sim N(0, I_{d_\theta + d_\beta})$ and V is a diagonal matrix defined later. Let $c_\omega(V^2, 1-\alpha)$ be the $1-\alpha$ quantile of $Z'V^2Z$ and let $\hat{V}_n$ be a consistent estimator of V (also defined later). Vuong (1989) uses the plug-in critical value $c_\omega(\hat{V}_n^2, 1-\alpha)$ for the pretest. To sum up, the two-step Vuong test rejects H_0 if and only if $n\hat{\omega}_n^2 > c_\omega(\hat{V}_n^2, 1-\alpha)$ and $|\hat{T}_n| > z_{\alpha/2}$.

3 Problems with the Classical Vuong Tests

In this section, I first use a Monte Carlo example to motivate the local asymptotic analysis, then derive the local asymptotic distribution, and finally use the new approximation to study the problems with the classical Vuong tests.

Both classical Vuong tests rely on $N(0,1)$ to approximate the distribution of $\hat{T}_n$ when H_0 holds and $\omega_{P_0}^2 > 0$. The following example illustrates that the actual distribution of $\hat{T}_n$ can be very different from $N(0,1)$ even when H_0 holds and $\omega_{P_0}^2 > 0$. As a result of the mismatch between the actual distribution and the pointwise asymptotic distribution, both classical Vuong tests demonstrate noticeable over-rejection. The multiple regression models are used to ensure that LR_{P_0} and $\omega_{P_0}^2$ have closed-form solutions under some DGPs, so that H_0 can be easily imposed and $\omega_{P_0}^2$ be flexibly adjusted. The particular DGPs are used to generate a clear mismatch between the distribution of $\hat{T}_n$ and $N(0,1)$.

Example 1 (Normal Regression)**.** *Suppose the two models to be compared are multiple regression mdoels with a known standard normal error term:*

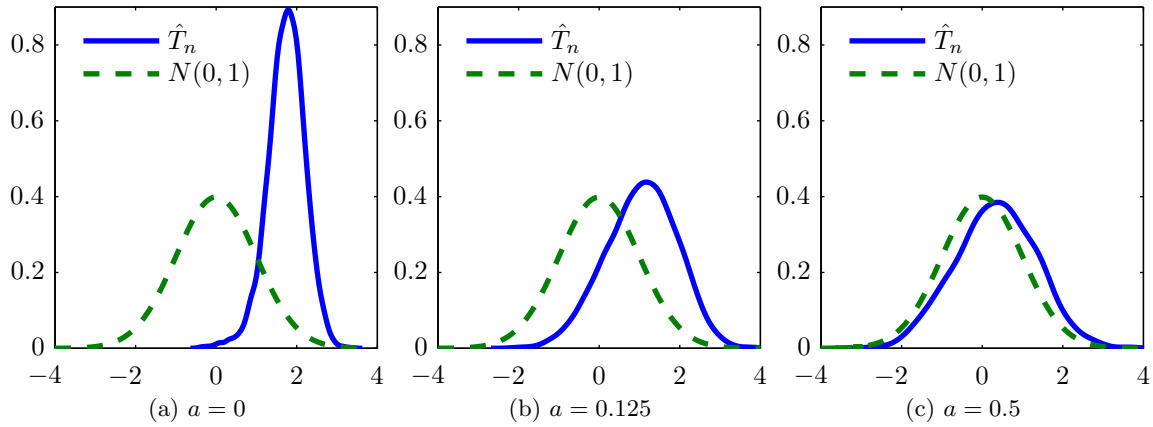
$$\begin{aligned} \mathcal{F}: & Y = \theta^{(0)} + \sum_{j=1}^{K_f} \theta^{(j)} Z_{f,j} + v, v|\vec{Z}_f, \vec{Z}_g \sim N(0,1), \\ \mathcal{G}: & Y = \beta^{(0)} + \sum_{j=1}^{K_g} \beta^{(j)} Z_{g,j} + u, u|\vec{Z}_f, \vec{Z}_g \sim N(0,1), \end{aligned} \tag{3.1}$$

where $\vec{Z}_f = (1, Z_{f,1}, \cdots, Z_{f,K_f})'$ *and* $\vec{Z}_g = (1, Z_{g,1}, \cdots, Z_{g,K_g})'$*. Consider DGPs of the form:*

$$Y = 1 + \sum_{j=1}^{K_f} Z_{f,j} + \sum_{j=1}^{K_g} Z_{g,j} + \varepsilon, \tag{3.2}$$

where $Z_{f,j} \sim N(0, a^2/K_f)$*,* $Z_{g,j} \sim N(0, a^2/K_g)$*,* $\varepsilon \sim N(0, 1-a^2)$*, and these variables are jointly independent. Under this DGP,* H_0 *holds for all* $a \in [0,1]$ *and* $\omega_{P_0}^2 = 2a^2 - a^4$*. Consider a large sample size* $n = 1000$*, a moderately large* $K_f = 15$ *and a small* $K_g = 1$*. The lengths of the regression equations,* K_f *and* K_g*, are designed to differ because the theory presented later shows that the higher-order bias in the log-likelihood ratio statistic is related to the relative magnitude of* K_f *and* K_g *(see Remark* (b) *of Theorem* 3.2*).*

The probability density functions of $\hat{T}_n$ *for three* a*'s are drawn in Figure* 1 *below along with the pdf of* $N(0,1)$*. The figure shows that the distribution of* $\hat{T}_n$ *for smaller* a*'s* (0 *and* 0.125) *is quite different from* $N(0,1)$*. As a result, the one-step Vuong test of nominal level* 5% *rejects* H_0 *with probability* 32% *when* $a = 0$ *and* 15% *when* $a = 0.125$*. Remarkably, the two-step Vuong test*



(a) $a = 0$ (b) $a = 0.125$ (c) $a = 0.5$

Figure 1: Pdfs of $\hat{T}_n$ versus that of $N(0,1)$ under different DGPs

also has a 15% *rejection rate when* $a = 0.125$*, suggesting that the rejection of the pretest is not sufficient indication that* $N(0,1)$ *well approximates the distribution of* $\hat{T}_n$*.*

The example shows that the pointwise-asymptotic distribution does not approximate the actual distribution of $\hat{T}_n$ well. This suggests that we should seek better approximation in a different asymptotic framework. I use the local asymptotic framework for this purpose. The local asymptotic framework is commonly used to study local power, weak identification, near unit root and boundary issues. It is suitable here because we encounter a boundary issue at $\omega_P^2 = 0$.

3.1 Local Asymptotic Theory

The pointwise asymptotic theory considers the limiting behavior of $\hat{T}_n$ under a fixed DGP, P. When $\omega_P^2 > 0$, as the sample size n increases to infinity, it becomes easier and easier to distinguish the positive ω_P^2 from zero. However, for a given finite sample, a positive ω_P^2 can be difficult to distinguish from zero due to sample variance. To reflect this uncertainty about the value of ω_P^2, I consider the sequence of DGPs, $\{P_n\}_{n=1}^{\infty}$, where $n\omega_{P_n}^2$ is approximately constant, in particular, $n\omega_{P_n}^2 \to \sigma^2 \in [0,\infty]$ as $n \to \infty$. The rate n is chosen so that the asymptotic distribution under $\{P_n\}_{n=1}^{\infty}$ varies smoothly with σ^2 over a range of distributions

To start, I impose the smoothness and compactness assumption on the models.

Assumption 3.1. (a) $\log f(y|z;\theta)$ *and* $\log g(y|z;\beta)$ *are three times continuously differentiable in* θ *and* β*, respectively, for all* $(y,z) \in \mathcal{X}$*, and*

(b) Θ *and* $\mathcal{B}$ *are compact.*

The assumptions on the DGPs are imposed in the definition of the maintained hypothesis below. To introduce the definition, I need some new notation. Let the population log-likelihood

functions be

$$ll_{f,P}(\theta) = E_P \log f_i(\theta) \text{ and } ll_{g,P}(\beta) = E_P \log g_i(\beta). \tag{3.3}$$

Let the expectation of the second derivative and that of the outer product of the first derivative of the log-density ratio be

$$A_P(\phi) = E_P[\partial^2 \Lambda_i(\phi)/\partial\phi\partial\phi'] \text{ and } B_P(\phi) = E_P[(\partial\Lambda_i(\phi)/\partial\phi)(\partial\Lambda_i(\phi)/\partial\phi')]. \tag{3.4}$$

These matrix-valued functions will appear in the local asymptotic distribution of $\hat{T}_n$.

Let $N_a(b)$ stand for an open ball of radius a around the point b. For a symmetric square matrix A, let $|eig|_{\min}(A)$ denote the minimum absolute value of the eigenvalues of A. Definition 3.1 below defines the maintained hypothesis, while Definition 3.2 defines the null hypothesis. The conditions in Definition 3.1 are the uniform version of their counterparts in Vuong (1989).

Definition 3.1. *For positive constants δ, M and function $\delta(\epsilon) : (0,\infty) \to (0,\infty)$, let $\mathcal{P}$ be the set of probability measures, P, on $\mathcal{X}$ such that*

(i) *the unique-identifiability condition holds, that is, $\exists\ \phi_P^* \in \Theta \times \mathcal{B}$, such that $\forall \epsilon > 0$,*

$$\sup_{\theta \in \Theta \setminus N_\epsilon(\theta_P^*)} ll_{f,P}(\theta) < ll_{f,P}(\theta_P^*) - \delta(\epsilon) \text{ and } \sup_{\beta \in \mathcal{B} \setminus N_\epsilon(\beta_P^*)} ll_{g,P}(\beta) < ll_{g,P}(\beta_P^*) - \delta(\epsilon)\ ,$$

(ii) *the pseudo-true parameters lie uniformly in the interior of their space: $N_\delta(\theta_P^*) \subseteq \Theta$ and $N_\delta(\beta_P^*) \subseteq \mathcal{B}$,*

(iii) $E_P \sup_{\theta\in\Theta} \left(|\log f_i(\theta)| + \|\frac{\partial \log f_i(\theta)}{\partial\theta}\| + \|\frac{\partial^2 \log f_i(\theta)}{\partial\theta\partial\theta'}\| + \sum_{j=1}^{d_\theta} \|\frac{\partial^3 \log f_i(\theta)}{\partial\theta_j\partial\theta\partial\theta'}\| \right)^{2+\delta}$
$+E_P \sup_{\beta\in\mathcal{B}} \left(|\log g_i(\beta)| + \|\frac{\partial \log g_i(\beta)}{\partial\beta}\| + \|\frac{\partial^2 \log g_i(\beta)}{\partial\beta\partial\beta'}\| + \sum_{j=1}^{d_\beta} \|\frac{\partial^3 \log g_i(\beta)}{\partial\beta_j\partial\beta\partial\beta'}\| \right)^{2+\delta} \leq M,$

(iv) $E_P[(\Lambda_i(\phi_P^*) - E_P\Lambda_i(\phi_P^*))/\omega_P(\phi_P^*)]^{2+\delta} \leq M$ *if* $\omega_P(\phi_P^*) > 0$, *and*

(v) *the Hessians are bounded away from singularity:* $\inf_{\phi\in N_\delta(\phi_P^*)} |eig|_{\min}(A_P(\phi)) \geq \delta$.

(vi) $(X_1, \cdots, X_n)$ *is an i.i.d. sample drawn from P.*

Definition 3.2. *Let* $\mathcal{P}_0 = \{P \in \mathcal{P} : E_P\Lambda_i(\phi_P^*) = 0\}$.

The correlation between the log-density ratio and its first derivatives will also appear in the local asymptotic distribution of $\hat{T}_n$. Thus we give it the notation:

$$\rho_P(\phi) = \omega_P^+(\phi) \left(D_P^{1/2}(\phi) \right)^+ E_P\{[\Lambda_i(\phi)][\partial\Lambda_i(\phi)/\partial\phi]\}. \tag{3.5}$$

where D_P is the diagonal matrix with the same diagonal as B_P ($D_P = Diag(B_P)$), $\omega_P^+(\phi) = \begin{cases} 0 & \text{if } \omega_P(\phi) = 0 \\ \omega_P^{-1}(\phi) & \text{if } \omega_P(\phi) \neq 0 \end{cases}$, and $\left(D_P^{1/2}(\phi) \right)^+$ is the Moore-Penrose inverse of the square root of

$D_P(\phi)$.[4] Now I define the sequences of DGPs along which I derive the asymptotic distribution of $\hat{T}_n$.

Definition 3.3. *Let $Seq(\sigma^2, A, B, \rho)$ be the set of sequences $\{P_n\}_{n=1}^{\infty}$ such that $P_n \in \mathcal{P}$ for every n, $n\omega_{P_n}^2 \to \sigma^2$, $A_{P_n}(\phi_{P_n}^*) \to A$, $B_{P_n}(\phi_{P_n}^*) \to B$ and $\rho_{P_n}(\phi_{P_n}^*) \to \rho$, as $n \to \infty$, where $\sigma \in [0, \infty]$, A is a block diagonal matrix with the upper-left block being $d_\theta \times d_\theta$ and negative semi-definite and the lower-right block being $d_\beta \times d_\beta$ and positive semi-definite, B is $(d_\theta + d_\beta) \times (d_\theta + d_\beta)$ and positive semi-definite, and ρ is a $d_\theta + d_\beta$ vector of correlation coefficients.*

Remark. Notice that σ is allowed to take values in the extended half real space $[0, \infty] \equiv R_+ \cup \{\infty\}$. The point ∞ is important for covering the cases where $\omega_{P_n}^2$ does not converge to zero or converges slowly to zero.

Let $D = Diag(B)$ and let $B^{1/2}$ denote the unique symmetric positive semi-definite matrix square root of B. Let $eig(A)$ denote the diagonal matrix formed by the eigenvalues of A.[5] The following theorem establishes the asymptotic distributions of $n^{1/2}\widehat{LR}_n$ and $n\hat{\omega}_n^2$ under each sequence in $Seq(\sigma^2, A, B, \rho)$.

Theorem 3.1. *Suppose Assumption 3.1 holds. Under a sequence $\{P_n\}_{n=1}^{\infty} \in Seq(\sigma^2, A, B, \rho)$ for some (σ^2, A, B, ρ) and $P_n \in \mathcal{P}_0$ for every n,*

(a) *if $\sigma \in [0, \infty)$, then*

$$\begin{pmatrix} n\widehat{LR}_n \\ n\hat{\omega}_n^2 \end{pmatrix} \to_d \begin{pmatrix} J_\Lambda \\ J_\omega \end{pmatrix} := \begin{pmatrix} \sigma Z_\Lambda - 2^{-1} Z_\phi' A^{-1} Z_\phi \\ \sigma^2 - 2\sigma\rho' D^{1/2} A^{-1} Z_\phi + Z_\phi' A^{-1} B A^{-1} Z_\phi \end{pmatrix},$$

where

$$\begin{pmatrix} Z_\Lambda \\ Z_\phi \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho' D^{1/2} \\ D^{1/2}\rho & B \end{pmatrix}\right),$$

(b) *if $\sigma = \infty$, then*

$$\omega_{P_n}^{-1} n^{1/2} \widehat{LR}_n \to_d N(0,1) \text{ and } \omega_{P_n}^{-2} \hat{\omega}_n^2 \to_p 1, \text{ and}$$

(c) $(J_\Lambda, J_\omega)'$ *can be rewritten as*

$$\begin{pmatrix} J_\Lambda(\sigma, \rho^*, V) \\ J_\omega(\sigma, \rho^*, V) \end{pmatrix} = \begin{pmatrix} \sigma Z_\Lambda - 2^{-1} Z_\phi^{*\prime} V Z_\phi^* \\ \sigma^2 - 2\sigma\rho^{*\prime} V Z_\phi^* + Z_\phi^{*\prime} V^2 Z_\phi^* \end{pmatrix}, \text{ where}$$

$V = eig(B^{1/2} A^{-1} B^{1/2})$, $(Z_\Lambda, Z_\phi^{*\prime})' \sim N(0, [1, \rho^{*\prime}; \rho^*, I_{d_\theta + d_\beta}])$ *and ρ^* is a solution to the equation $B^{1/2} Q \rho^* = D^{1/2}\rho$, where Q is an orthonormal matrix that satisfies: $QVQ' = B^{1/2} A B^{1/2}$.*

[4] We allow D_P to be singular to facilitate the extension to moment-based models in Section 6. In moment-based models, the corresponding D_P matrices are always singular if one of the models is correctly specified.

[5] The order in which each eigenvalue appears is not important.

Remark. Note that under the hypothesis $H_{00} : \omega_P^2 = 0$, we have $\sigma = 0$. Part (c) implies $n\hat{\omega}_n^2 \to_d Z_\phi^{*\prime} V^2 Z_\phi^*$, which is consistent with Vuong's (1989) derivation reviewed in (2.8). The matrix V can be consistently estimated by the plug-in estimator $\hat{V}_n = eig(\hat{B}_n^{1/2} \hat{A}_n^{-1} \hat{B}_n^{1/2})$ where $\hat{A}_n = \hat{A}_n(\hat{\phi}_n)$ and $\hat{B}_n = \hat{B}_n(\hat{\phi}_n)$ with

$$\hat{A}_n(\phi) = n^{-1} \sum_{i=1}^n \frac{\partial^2 \Lambda_i(\phi)}{\partial \phi \partial \phi'} \text{ and } \hat{B}_n(\phi) = n^{-1} \sum_{i=1}^n \frac{\partial \Lambda_i(\phi)}{\partial \phi} \frac{\partial \Lambda_i(\phi)}{\partial \phi'}. \tag{3.6}$$

This $\hat{V}_n$ is the one in the pretest critical value $c_\omega(\hat{V}_n^2, 1-\alpha)$ of the two-step Voung test.

The theorem gives the local asymptotic distribution of $\hat{T}_n = \sqrt{n}\widehat{LR}_n / \hat{\omega}_n$:

$$\hat{T}_n \to_d J(\sigma, \rho^*, V) := \frac{J_\Lambda(\sigma, \rho^*, V)}{J_\omega^{1/2}(\sigma, \rho^*, V)} = \frac{\sigma Z_\Lambda - 2^{-1} Z_\phi^{*\prime} V Z_\phi^*}{\sqrt{\sigma^2 - 2\sigma \rho^{*\prime} V Z_\phi^* + Z_\phi^{*\prime} V^2 Z_\phi^*}}. \tag{3.7}$$

This asymptotic distribution shows that $\hat{T}_n$ can be distributed quite differently from $N(0,1)$ under H_0. Next I would like to show the extent to which the quantile of the different distribution can be bigger than the standard normal quantiles because that determines the size distortion of the classical Vuong tests. For this purpose, I examine the new asymptotic distribution closely in the next two subsections.

3.2 Bias in the Numerator

The random variable on the right-hand side (r.h.s.) of (3.7) is complicated because it is a ratio of two functions of generalized chi-squared random variables and both functions depend on the unknown parameters (σ, ρ^*, V). To study its distribution, let us first focus on the numerator.

The numerator is the local asymptotic limit of $n\widehat{LR}_n$, which is a sample-analogue estimator of nLR_{P_n}. In the local aymptotic framework of Theorem 3.1, we see that the estimator $n\widehat{LR}_n$ is a biased estimator because when $nLR_{P_n} = 0$ (i.e., $P_n \in \mathcal{P}_0$), $E[J_\Lambda(\sigma, \rho^*, V)] = -trace(V)/2$, which typically is nonzero. If we treat the denominator as deterministic for a moment, we see that the bias in $n\widehat{LR}_n$ causes a nonzero mean in the null distribution of $\hat{T}_n \equiv n\widehat{LR}_n / \sqrt{n\hat{\omega}_n^2}$. This is exactly the situation depicted in Figure 1. In some cases $-trace(V)/2$ can be very large in magnitude, making the two curves in Figure 1 very far apart and making the null rejection probability of the classical Vuong tests arbitrarily close to one. Theorem 3.2 below illustrates such extreme size distortion.[6] In the theorem, "$tr(A)$" stands for the trace of the matrix A.

[6]Notice that Theorem 3.2 uses sequential asymptotics, where I let n go to infinity and then let k go to infinity. Because of this, the theorem does not describe the situation where k is almost as large as n; rather, it describes the situation where k is moderately large, but n is much larger. With additional effort, the same statement can be shown to hold with k and n going to infinity simultaneously but k at some slower rate than n. But since the additional effort does not add new insights to the discussion, it is omitted for brevity.

Theorem 3.2. *For $k = d_\theta + d_\beta$, let the sequence (σ_k, V_k) be such that, $\sqrt{tr(V_k^4)}/tr(V_k^2) \to 0$, $\sigma_k/\sqrt{tr(V_k^2)} \to \infty$, and $tr(V_k)/\sigma_k \to -\infty$ as $k \to \infty$. For each k, let $\{P_{n,k} \in \mathcal{P}_0\}_{n=1}^{\infty} \in Seq(\sigma_k^2, A_k, B_k, \rho_k)$ for some $\left(\sigma_k^2, A_k, B_k, \rho_k\right)$ satisfying $V_k = eig\left(B_k^{1/2} A_k^{-1} B_k^{1/2}\right)$. Suppose Assumption* 3.1 *holds. Then* (a) *the size of the two-step classical Vuong test size approaches* 100%*:*

$$\lim_{k\to\infty} \lim_{n\to\infty} \Pr_{P_{n,k}} \left(n\hat{\omega}_n^2 > c_\omega(\hat{V}_n^2, 1-\alpha) \;\&\; n^{1/2}\widehat{LR}_n/\hat{\omega}_n > z_{\alpha/2} \right) = 1,$$

and (b) *the size of the one-step classical Vuong test approaches* 100%*:*

$$\lim_{k\to\infty} \lim_{n\to\infty} \Pr_{P_{n,k}} \left(n^{1/2}\widehat{LR}_n/\hat{\omega}_n > z_{\alpha/2} \right) = 1,$$

Remark. (a) In finite samples, the theorem describes the situation in which $n\omega_P^2$ is small relative to $|tr\,(V)|$, which arises when the best-fitting pdfs of the two models are close for a given $|tr\,(V)|$. This may occur either when the models are overlapping or when they are strictly nonnested but have elements that are close. Therefore, the extreme size distortion is possible even for strictly nonnested models.

(b) The condition $tr(V_k)/\sigma_k \to -\infty$ is important for the rejection probability to go to 1 as $k \to \infty$. Observe that $tr\,(V) = tr\left(A^{-1}B\right) = tr\left(A_f^{-1}B_f\right) - tr\left(A_g^{-1}B_g\right)$, where A_f, B_f are the limits of the Hessian and the gradient version of the Fisher information matrices of model $\mathcal{F}$ and A_g, B_g are those of model $\mathcal{G}$. For models for which the information identify holds or approximately holds, we have $tr\,(V) \approx -d_\theta + d_\beta$. That is, for such models, the extreme over-rejection occurs when one model is much less parsimonious than the other. And the over-rejecting classical Vuong tests are biased in favor of the more complex model. In this sense, the classical Vuong tests reward model complexity. However, it is worth noting that in the Vuong null hypothesis, H_0, the models are compared purely based on the KL distance and model complexity is neither rewarded nor penalized. This impartiality is fully respected by the new nondegenerate test that I propose below.

3.3 Random Denominator

The previous subsection treats the denominator $J_\omega^{1/2}(\sigma, \rho^*, V)$ as if it is deterministic, but, in fact, it is random, especially when k is small. Rewrite $J_\omega(\sigma, \rho^*, V)$ as

$$J_\omega(\sigma, \rho^*, V) = \sigma^2(1 - \rho^{*\prime}\rho^*) + (VZ_\phi^* - \sigma\rho^*)'(VZ_\phi^* - \sigma\rho^*). \tag{3.8}$$

Although it is clear that $J_\omega(\sigma, \rho^*, V)$ is always nonnegative, it can take values close to zero with significant probability when $(1 - \rho^{*\prime}\rho^*)$ is small and at the same time $\sigma V^+\rho^*$ lies in an area where the probability density of Z_ϕ^* is high. If it takes values close to zero at the time that $J_\Lambda(\sigma, \rho^*, V)$ is non-zero, the ratio $J_\Lambda(\sigma, \rho^*, V)/J_\omega^{1/2}(\sigma, \rho^*, V)$ can be large, creating a fat tail for

its distribution.

It is difficult to study the fat tail feature of $J_\Lambda(\sigma,\rho^*,V)/J_\omega^{1/2}(\sigma,\rho^*,V)$ for general σ,ρ^* and V. Here, I study $J_\Lambda(\sigma,\rho^*,V)/J_\omega^{1/2}(\sigma,\rho^*,V)$ for a specific choice of σ,ρ^* and V. I plot its pdf by simulating the normal random vectors Z_ϕ^* and Z_Λ and compute its tail probabilities using numerical tools. The parameters I consider are $\rho^*=(1,0)$ and $V=diag((1,0)')$ and $\sigma=1.5$.[7] Then simple algebra shows

$$\frac{J_\Lambda(\sigma,\rho^*,V)}{J_\omega^{1/2}(\sigma,\rho^*,V)}=\frac{3Z_\Lambda-Z_\Lambda^2}{2|1.5-Z_\Lambda|}, \tag{3.9}$$

where $Z_\Lambda\sim N(0,1)$. In this specification, the denominator is close to zero and the numerator is a large positive constant when Z_Λ is close to 1.5, and the probability density of Z_Λ at 1.5 is not very small. Thus, we can expect a fat tail in the distribution of the ratio. This is confirmed in Figure 2, which shows the pdf of the ratio.

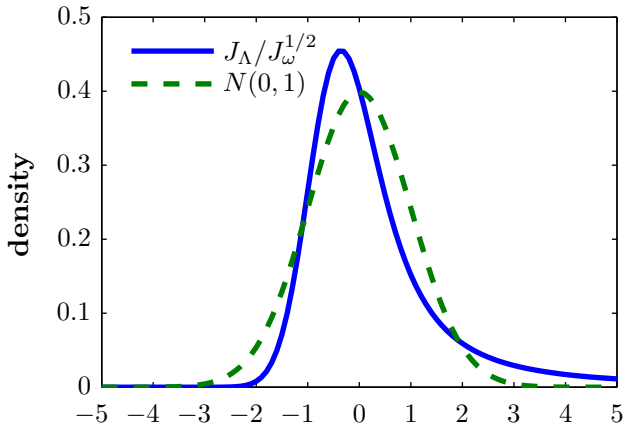


Figure 2: Fat-tailed local asymptotic distribution against $N(0,1)$

Numerical computation yields $\Pr\left(J_\omega^{-1/2}J_\Lambda>z_{0.05/2}\right)\approx 14\%$. This suggests that a nominal level 5% one-step Vuong test can pick one model with probability close to 14% when applied to models with ρ^* and $V/\sqrt{tr(V^2)}$ close to $(1,0,...,0)$ and $diag((1,0,...,0)')$, respectively, while this probability should only be 2.5% according to the pointwise asymptotic theory. [8]

4 New Nondegenerate Test

The previous section shows that the null distribution of $\hat{T}_n$ can be rather different from $N(0,1)$ because of an asymptotic bias in the numerator $n\widehat{LR}_n$ and because of the randomness in the

[7] This set of parameters is motivated by the discussion in the above paragraph and by simulation experiments. By simulating the quantiles of the ratio with many sets of parameters, I find that this set yields the worst (largest) tail probability for the ratio $J_\Lambda(\sigma,\rho^*,V)/J_\omega^{1/2}(\sigma,\rho^*,V)$.

[8] On the other hand, in this specification, the two-step Vuong test has no over-rejection because the variance test tends not to reject when $\left|J_\omega^{-1/2}J_\Lambda\right|$ has a large quantile. However, using the variance test to control the size results in large power loss, as illustrated in Example 2 in Section 7 below.

denominator $n^{1/2}\hat{\omega}_n$. Here I propose corrections to address both issues. The corrected test statistic combined with a simulated critical value forms my new nondegenerate Vuong test.

4.1 The New Test and Its Asymptotic Size

The first correction is the bias correction to the numerator. Section 3.2 shows that $\widehat{LR}_n$ as an estimator of $E_{P_0}\Lambda_i(\phi^*_{P_0})$ has a $O(1/n)$ bias $tr\left(V\right)/2n$. I use the consistent estimator $\hat{V}_n$ of V to correct the bias. The bias corrected numerator is defined as:

$$\widehat{LR}_n^{\text{mod}} = \widehat{LR}_n + tr\left(\hat{V}_n\right)/(2n), \tag{4.1}$$

where "mod" stands for "modified."

The second correction is the adjustment to the denominator. In Section 3.3, I show that the random denominator can take values close to zero with noneglibible probablity, creating a fat tail for $\hat{T}_n$. I counteract the effect of the randomness by adding a positive constant term to it

$$\left(\hat{\omega}_n^{\text{mod}}(c)\right)^2 = \hat{\omega}_n^2 + c \cdot tr(\hat{V}_n^2)/n, \tag{4.2}$$

where c is a positive constant. We discuss a data-dependent method to choose c below.

Then, the new test statistic is

$$\hat{T}_n^{\text{mod}}(c) = \frac{n^{1/2}\widehat{LR}_n^{\text{mod}}}{\hat{\omega}_n^{\text{mod}}(c)}. \tag{4.3}$$

Under the DGP sequence in Theorem 3.1, one can show that

$$\hat{T}_n^{\text{mod}}(c) \to_d J(\sigma, \rho^*, V, c) := \frac{J_\Lambda(\sigma, \rho^*, V) + tr(V)/2}{\sqrt{J_\omega(\sigma, \rho^*, V) + c \cdot tr(V^2)}}. \tag{4.4}$$

The random variable on the r.h.s. has tails that are close to those of $N(0,1)$, but still not the same as those of $N(0,1)$. Thus, if we use the $N(0,1)$ critical values, the new test might still have some size distortion, even though not as bad as the classical Vuong tests.

To completely remove the asymptotic size distortion, I propose a simulation-based critical value: $cv(1-\alpha, \hat{V}_n, c)$, where

$$cv(1-\alpha, V, c) = \sup_{\sigma \in [0,\infty], \rho^* : \|\rho^*\| \leq 1} F^{-1}_{|J(\sigma,\rho^*,V,c)|}(1-\alpha), \tag{4.5}$$

where $F_X^{-1}(\tau)$ denotes the τ-quantile of X. The reason that we take the supremum over σ and ρ^* is that these parameters cannot be consistently estimated.[9] The supremum gives us a consistent

[9]The parameter σ^2 cannot be consistently estimated because it is the limit of $n\omega^2_{P_n}$ and $\omega^2_{P_n}$ cannot be estimated n^{-1}-consistently. The parameter ρ^* cannot be estimated because $\rho_{P_n}(\phi^*_{P_n})$ cannot be consistently estimated when $\omega^2_{P_n}$ drifts to zero.

upper bound for the quantile of $J(\sigma, \rho^*, V, c)$. Note that the critical value is weakly bigger than $z_{\alpha/2}$ because $J(\infty, \rho^*, V, c) \sim N(0,1)$ for any (ρ^*, V, c).

With the modified test statistic and the critical value, the new nondegenerate test rejects H_0 if $|\hat{T}_n^{\text{mod}}(c)| > cv(1-\alpha, \hat{V}_n, c)$. Theorem 4.1 below shows that this test has well-controlled asymptotic size.

Theorem 4.1. *Suppose Assumption* 3.1 *holds. Then, for any* $c \geq 0$*,*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} \Pr_P \left(|\hat{T}_n^{\text{mod}}(c)| > cv(1-\alpha, \hat{V}_n, c) \right) \leq \alpha.$$

4.2 Implementation of the Test

I provide MATLAB and STATA codes that implement the new nondegenerate test. The codes implement the following steps, which summarize the procedure introduced above.

- Step 1. Compute the maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\beta}_n$. From the maximum likelihood procedure, also obtain the Hessian matrices of the two models respectively, denoted $\hat{A}_{n,f}$ and $\hat{A}_{n,g}$, as well as the first derivatives of the log-densities: $\{\partial \log f_i(\hat{\theta}_n)/\partial\theta\}_{i=1}^n$ and $\{\partial \log g_i(\hat{\beta}_n)/\partial\beta\}_{i=1}^n$.
- Step 2. Let $\hat{A}_n = \begin{pmatrix} \hat{A}_{n,f} & 0 \\ 0 & -\hat{A}_{n,g} \end{pmatrix}$ and let $\hat{B}_n$ be the sample covariance matrix of the vector $[\partial \log f_i(\hat{\theta}_n)/\partial\theta', \partial \log g_i(\hat{\beta}_n)/\partial\beta']'$. Let $\hat{V}_n = eig(\hat{B}_n^{1/2}\hat{A}_n^{-1}\hat{B}_n^{1/2})$.
- Step 3. Choose a c (see the choice of c below), and obtain $\hat{T}_n^{\text{mod}}(c)$ according to (4.3).
- Step 4. Draw an independent sample $\{(Z_{\Lambda,s}, Z_{\phi,s}^{*,\prime})'\}_{s=1}^S$ from $N(0, [1, \rho^{*,\prime}; \rho^{*,\prime}, \hat{V}_n])$ for a large number S.
- Step 5. For a fixed σ, ρ^*, and $\hat{V}_n$ obtained in Step 2 and c chosen in Step 3,
  - for each s, plug $(Z_{\Lambda,s}, Z_{\phi,s}^{*,\prime})'$ into (4.4) to obtain $J_s := J_s(\sigma, \rho^*, V, c)$, and then
  - let $F^{-1}_{|J(\sigma,\rho^*,\hat{V}_n,c)|}(1-\alpha)$ be the $1-\alpha$ quantile of the sample $\{J_1, \cdots, J_S\}$.
- Step 6. Use the recommended procedure for computing the supremum below to obtain $cv(1-\alpha, \hat{V}_n, c)$. Make sure *not* to redo Step 4 when an $F^{-1}_{|J(\sigma,\rho^*,\hat{V}_n,c)|}(1-\alpha)$ is computed for each (σ, ρ^*, c).
- Step 7. Compare $\hat{T}_n^{\text{mod}}(c)$ to $cv(1-\alpha, \hat{V}_n, c)$ and
  - reject H_0 and pick model $\mathcal{F}$ if $\hat{T}_n^{\text{mod}}(c) > cv(1-\alpha, \hat{V}_n, c)$,
  - reject H_0 and pick model $\mathcal{G}$ if $\hat{T}_n^{\text{mod}}(c) < -cv(1-\alpha, \hat{V}_n, c)$, and
  - do not reject H_0 otherwise.

Next, I discuss two important implementation details of the new test: the choice of c and the computation of the supremum. I recommend detailed procedures for both based on experience from the Monte Carlo experiments and the empirical application. I discuss the general case first and then discuss the nested case in which things simplify.

Choice of c. The constant c trades off the power of the test in different areas of the DGP space. Larger c leads to smaller $cv(1-\alpha, \hat{V}_n, c)$, and thus increases the power of the test against the alternatives under which $\omega^2_{P_0}$ is large and decreases the power against alternatives under which $\omega^2_{P_0}$ is small. I propose a data-dependent procedure that balances the power across the DGP space. The procedure is as follows:

- Start with $c = 0$ and obtain $cv(1-\alpha, \hat{V}_n, 0)$. If $cv(1-\alpha, \hat{V}_n, 0) - z_{\alpha/2}$ is smaller than a tolerance level, say 0.1, stop here and use $c = 0$. Otherwise, proceed to the next step.

- Increase c until it reaches a point c^* such that $cv(1-\alpha, \hat{V}_n, c^*) - z_{\alpha/2} \approx 0.1$. Use $c = c^*$. To find this c^*, one can either use a grid search, or a simplex-based method. Finding c^* numerically is easy due to the monotonicity of $cv(1-\alpha, \hat{V}_n, \cdot)$.

Computation of the Supremum. The critical value $cv(1-\alpha, \hat{V}_n, c)$ is defined as a supremum over a $k+1$ dimensional parameter, which can be time-consuming to compute. My experience with the Monte Carlo examples suggest two simplifications.

- In my experience, the supremum over ρ^* is always achieved at extremum points of the space $\{\rho^* \geq 0 : \|\rho^*\| \leq 1\}$. In particular, it is always achieved at $\rho^* = (0, \cdots, 0, 1, 0, \cdots, 0)'$, where the value 1 occurs at the same location as the largest element (in absolute value) of $\hat{V}_n$ on the diagonal of $\hat{V}_n$. For example, if $diag(\hat{V}_n) = (-.99, 0.1, .5)'$, then the supremum over ρ^* is always achieved at $\rho^* = (1, 0, 0)'$. Thus I recommend focusing on this single point of ρ^*.

- Not all points on the space, $[0, \infty]$, of σ are equally relevant for the supremum. In my experience, the supremum, if bigger than $z_{\alpha/2}$, typically occurs at the σ such that $\sigma/\sqrt{tr(V^2)} \in [0, 5]$. Thus, I recommend searching for more points in this interval than outside. If an optimization routine that requires a starting value is used, I recommend picking starting values from this interval. Sometimes, the supremum may be achieved at $\sigma = \infty$. That is typically the case where the supremum is smaller than $z_{\alpha/2}$ up to the simulation error. In that case, when to stop the search should not make a big difference. Furthermore, I recommend always considering $\sigma = \infty$, at which point $F^{-1}_{|J(\sigma,\rho^*,\hat{V}_n,c)|}(1-\alpha) = z_{\alpha/2}$.

Nested Case. If we know that $\mathcal{F}$ and $\mathcal{G}$ are nested, then σ can only take the value 0. Moreover, in this case, a one-sided test is more useful than a two-sided one because the nested model cannot be strictly closer to the truth than the nesting one according to the Kullback-Leibler distance. Suppose without loss of generality that $\mathcal{G} \subset \mathcal{F}$. Then, typically, one would like to test

$H_0 : LR_{P_0} = 0$ versus $H_1 : LR_{P_0} > 0$. Then, the test statistic can stay unchanged, while the critical value should be changed to

$$cv^{nested}(1-\alpha, V, c) = F^{-1}_{J(0,-,V,c)}(1-\alpha), \tag{4.6}$$

where the "$-$" replaces the ρ^* argument in $J(0,-,V,c)$ because when $\sigma = 0$, that argument is redundant. The test rejects H_0 in favor of H_1 if $\hat{T}_n^{\text{mod}}(c) > cv^{nested}(1-\alpha, V, c)$. Because nondegeneracy does not occur in the nested case, making the critical value small is not as important as in the nonnested case. It is thus reasonable to simply set $c = 0$. This test is a misspecification-robust alternative to the usual χ^2 test for nested models; that is, this test allows the nesting model ($\mathcal{F}$) to be misspecified.

5 Power Properties of the Nondegenerate Vuong Test

The new nondegenerate Vuong test has power approaching one against all fixed alternatives for the same reason that the classical Vuong tests do. This is because the modifications to both $\widehat{LR}_n$ and $\hat{\omega}_n^2$ are of smaller order than $\widehat{LR}_n$ and $\hat{\omega}_n^2$ under fixed alternatives and thus vanish in the limit. The consistency of the classical Vuong tests is shown in Theorems 5.1 and 6.3 in Vuong (1989). I do not repeat the work but rather focus on the local power of the new test.

Two types of local alternative sequences are defined by Assumption 5.1 and Assumption 5.2, respectively. The former defines the $n^{-1/2}$-local alternatives, whereas the latter defines the local alternatives under which $E_{P_n}\Lambda_i(\phi^*_{P_n})$ converges to zero at a rate faster than $n^{-1/2}$.

Assumption 5.1. *The sequence of true DGPs* $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$ *weakly converges to a* $P_0 \in \mathcal{P}_0$*, and* (a) $n^{1/2}LR_{P_n} \to d \in R/\{0\}$*,* (b) $\omega^2_{P_n} \to \omega^2_{P_0} \in [0,\infty)$*, and* (c) $V_{P_n} := eig(B^{1/2}_{P_n}A^{-1}_{P_n}B^{1/2}_{P_n}) \to V_{P_0}$*.*

Assumption 5.2. *The sequence of true DGPs* $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$ *weakly converges to a* $P_0 \in \mathcal{P}_0$*, and* (a) $nLR_{P_n} \to \infty$*, and* (b) $n\omega^2_{P_n} \to \sigma^2_{\infty} \in [0,\infty)$*.*

Theorem 5.1 below shows that the modified test has nontrivial power against the above-defined local alternatives.

Theorem 5.1. (a) *Under Assumptions* 3.1 *and* 5.1*, for any* $c > 0$*, ,*

$$\lim_{n\to\infty} \Pr{}_{P_n}\left(\hat{T}_n^{\text{mod}}(c) > cv(1-\alpha, \hat{V}_n, c)\right) = \Phi\left(d/\omega_{P_0} - cv(1-\alpha, V_{P_0}, c)\right) \text{ and}$$
$$\lim_{n\to\infty} \Pr{}_{P_n}\left(\hat{T}_n^{\text{mod}}(c) < -cv(1-\alpha, \hat{V}_n, c)\right) = \Phi\left(-d/\omega_{P_0} - cv(1-\alpha, V_{P_0}, c)\right).$$

(b) *Under Assumptions* 3.1 *and* 5.2*, for any* $c > 0$*,*

$$\lim_{n\to\infty} \Pr{}_{P_n}\left(\hat{T}_n^{\text{mod}}(c) > cv(1-\alpha, \hat{V}_n, c)\right) = 1, \text{ and}$$
$$\lim_{n\to\infty} \Pr{}_{P_n}\left(\hat{T}_n^{\text{mod}}(c) < -cv(1-\alpha, \hat{V}_n, c)\right) = 0.$$

Remark. (a) Part (a) shows the asymptotic power function of the new nondegenerate test when $\omega_{P_0}^2$ may or may not be small. This describes the power of the test to detect the better model when both candidate models may be globally misspecified. Notice that the $n^{-1/2}$ is the standard rate when one considers $E_P\Lambda_i(\phi_P^*)$ as a parameter and the test as one based on the t-statistic of this parameter (except with preestimated parameter $\hat{\phi}_n$).

(b) Part (b) shows the local power of the new nondegenerate test when $\omega_{P_0}^2$ is small. This describes the power of the test to detect the better model when the candidate models are correctly specified or mildly misspecified, which arguably are the more relevant situations in practice. The faster rate "n^{-1}" is rather encouraging as it implies that the new test has the ability to discern very small deviations from the null in these empirically relevant situations.

6 Moment Based Models

In all of the discussions above, the fact that $\log f_i(\theta)$ and $\log g_i(\beta)$ are logarithms of density functions is not essential. Neither is the fact that $\hat{\theta}_n$ and $\hat{\beta}_n$ are maximizers of log-likelihoods. Once the maximum likelihood structure is relaxed, the new test proposed above can be adapted to a much broader class of problems. In particular, it can be used to compare moment-based models by the generalized empirical likelihood (GEL, see Smith (1997) or Kitamura (2006)) criteria. I describe the extended framework in this section.

Still let $\mathcal{F}$ and $\mathcal{G}$ denote the two competing models, which are now moment-based models:

$$\begin{aligned}\mathcal{F} &= \left\{F : \int m_f(x,\psi_f)\,dF(x) = 0 \text{ for some } \psi_f \in \Psi_f \subset R^{d_{\psi_f}}\right\},\\ \mathcal{G} &= \left\{F : \int m_g(x,\psi_g)\,dF(x) = 0 \text{ for some } \psi_g \in \Psi_g \subset R^{d_{\psi_g}}\right\}.\end{aligned} \tag{6.1}$$

where m_f, m_g are moment functions known up to the parameters ψ_f and ψ_g. The GEL distance from the models to the true distribution P_0 can be written into the following general form:

$$d(\mathcal{F},P_0) = E_{P_0}\left[L_f\left(X_i,\psi_{f,P_0}^*,\gamma_{f,P_0}^*\right)\right] \text{ and } d(\mathcal{G},P_0) = E_{P_0}\left[L_g\left(X_i,\psi_{g,P_0}^*,\gamma_{g,P_0}^*\right)\right], \tag{6.2}$$

where (ignoring the subscripts f and g indexing the models) L is a function known up to the parameters $\theta := (\psi',\gamma')'$, γ is the Lagrange multiplier introduced to write the GEL distances into the form above, and $\theta_P^* := (\psi_{P_0}^{*\prime},\gamma_{P_0}^{*\prime})'$ denotes the pseudo-true value of the parameter under the data distribution P. Typically, θ_P^* is a saddle point of $E_P[L(X_i,\phi,\gamma)]$, that is,

$$\begin{aligned}\psi_P^* &= \arg\max_{\psi} E_P[L(X_i,\psi,\gamma_P^*(\psi))] \text{ and } \gamma_P^* = \gamma_P^*(\psi_P^*), \text{ where}\\ \gamma_P^*(\psi) &= \arg\min_{\gamma} E_P[L(X_i,\psi,\gamma)].\end{aligned} \tag{6.3}$$

The GEL distance includes the following well-known examples. Additional examples are given

in Kitamura (2006).

- Empirical Likelihood (EL): $L(x,\theta) = -\log(1-\gamma' m(x,\psi))$.
- Exponential Tilting (ET): $L(x,\theta) = \exp(\gamma' m(x,\psi))$.
- Continuous Updating GMM (CUE): $L(x,\theta) = (\gamma' m(x,\psi)+1)^2/2$.

The Vuong-type hypothesis for the moment models based on the GEL distance is

$$H_0 : d(\mathcal{F}, P_0) = d(\mathcal{G}, P_0). \tag{6.4}$$

Now let ϕ denote $(\theta_f', \theta_g')'$ and

$$\Lambda_i(\phi) = -L_f(X_i, \theta_f) + L_g(X_i, \theta_g); \tag{6.5}$$

then H_0 can be rewritten exactly as (2.2).

Let $\hat{\phi}_n$ be the GEL estimator of $\phi^*_{P_0}$, that is, $\hat{\phi}_n = (\hat{\psi}_{f,n}', \hat{\gamma}_{f,n}', \hat{\psi}_{g,n}', \hat{\gamma}_{g,n}')'$ where

$$\begin{aligned} \hat{\psi}_{j,n} &= \arg\max_{\psi_j} n^{-1}\sum_{i=1}^n L(X_i, \psi_j, \hat{\gamma}_{j,n}(\psi_j)) \text{ and } \hat{\gamma}_{j,n} = \hat{\gamma}_{j,n}(\hat{\psi}_{j,n}), \text{ where} \\ \hat{\gamma}_{j,n}(\psi_j) &= \arg\min_{\gamma_j} n^{-1}\sum_{i=1}^n L(X_i, \psi_j, \gamma_j),\ j \in \{f,g\}. \end{aligned} \tag{6.6}$$

Let $\widehat{LR}_n$, $\hat{\omega}_n^2$, $c_\omega(V, 1-\alpha)$, ω_P^2, $A_P(\phi)$, $B_P(\phi)$ and $\rho_P(\phi)$ be defined in the same way as in Sections 2-3 but with the new $\Lambda_i(\phi)$ and ϕ. Let $\hat{A}_n$, $\hat{B}_n$, $\hat{V}_n$, $\widehat{LR}_n^{\text{mod}}$, $\hat{\omega}_n^{\text{mod}}(c)$, $\hat{T}_n^{\text{mod}}(c)$ and $cv(1-\alpha, V, c)$ be defined as in Section 4 but with the new $\Lambda_i(\phi)$ and ϕ. Below I show that (i) both the one-step test and the two-step test analogous to those in Vuong (1989) have size distortion, and (ii) the new nondegenerate test based on $\hat{T}_n^{\text{mod}}(c)$ and $cv(1-\alpha, \hat{V}_n, c)$ is asymptotically uniformly valid and has good power properties.

To describe the results, we first introduce the assumptions. These assumptions are similar to the ones used in previous sections for the parametric models, but some are adapted to fit into the moment-based model framework. Let Θ_f and Θ_g be the parameter spaces of model $\mathcal{F}$ and $\mathcal{G}$ respectively. I first introduce the assumption on the models:

Assumption 6.1. (a) *$L_f(x,\theta_f)$ and $L_g(x,\theta_g)$ are three times continuously differentiable in θ_f and θ_g respectively, for all $x \in \mathcal{X}$, and*

(b) *Θ_f and Θ_g are compact.*

Next, I define the space of DGPs and the subspace that satisfies H_0.

Definition 6.1. *For positive constants δ and M, let $\mathcal{P}$ be the set of probability measures, P, on $\mathcal{X}$ such that*

(i) *the pseudo-true parameters satisfy the first-order conditions:* $\exists\ \phi_P^* \equiv (\theta_{f,P}^{*\prime}, \theta_{g,P}^{*\prime})' \in \Theta_f \times \Theta_g$, *such that*

$$0 = \frac{\partial E_P\left[L_f\left(X_i, \theta_{f,P}^*\right)\right]}{\partial \theta_f} = \frac{\partial E_P\left[L_g\left(X_i, \theta_{g,P}^*\right)\right]}{\partial \theta_g}, \tag{6.7}$$

(ii) *the pseudo-true parameters lie uniformly in the interior of their space:* $N_\delta(\theta_{f,P}^*) \subseteq \Theta_f$ *and* $N_\delta(\theta_{g,P}^*) \subseteq \Theta_g$,

(iii) $E_P \sup_{\theta_j \in \Theta_j} \left(|L_j(X_i, \theta_j)| + \| \frac{\partial L_j(X_i,\theta_j)}{\partial \theta_j} \| + \| \frac{\partial^2 L_j(X_i,\theta_j)}{\partial \theta_j \partial \theta_j'} \| + \sum_{r=1}^{d_{\theta_j}} \| \frac{\partial^3 L_j(X_i,\theta_j)}{\partial \theta_{j,r} \partial \theta_j \partial \theta_j'} \| \right)^{2+\delta} \leq M$ *for* $j \in \{f, g\}$,

(iv) $E_P[(\Lambda_i(\phi_P^*) - E_P\Lambda_i(\phi_P^*))/\omega_P(\phi_P^*)]^{2+\delta} \leq M$ *if* $\omega_P(\phi_P^*) > 0$,

(v) *the Hessians are bounded away from singularity:* $\inf_{\phi \in N_\delta(\phi_P^*)} |eig|_{\min}(A_P(\phi)) \geq \delta$, *and*

(vi) $(X_1, \cdots, X_n)$ *is an i.i.d. sample drawn from* P.

Definition 6.2. $\mathcal{P}_0 = \{P \in \mathcal{P} : E_P\Lambda_i(\phi_P^*) = 0\}$.

Because we replaced the unique-identifiability assumption – condition (i) of Definition 3.1 – by the first-order condition – condition (i) of Definition 6.1, we need to supplement it with the following additional assumption:

Assumption 6.2. (a) *Under the sequence* $\{P_n\}_{n=1}^\infty \in Seq(\sigma^2, A, B, \rho)$ *for some* (σ^2, A, B, ρ), *we have* $\|\hat{\phi}_n - \phi_{P_n}^*\| \to_p 0$, *and*

(b) *the finite sample first-order condition holds:* $0 = \sum_{i=1}^n \frac{\partial L_f\left(X_i, \hat{\theta}_{f,n}\right)}{\partial \theta_f} = \sum_{i=1}^n \frac{\partial L_g\left(X_i, \hat{\theta}_{g,n}\right)}{\partial \theta_g}$.

Remark. Assumption 6.2(a) assumes the consistency of the estimator $\hat{\phi}_n$ under drifting sequences of DGPs. For moment-based models, Kitamura (2000) (Lemma 1) and Chen, Hong, and Shum (2007) (Theorem 3) give primitive sufficient conditions for such consistency in the cases of exponential tilting (ET) and empirical likelihood (EL), respectively. Consistency for the other GEL distances can be established similarly.[10] Assumption 6.2(b) is satisfied for GEL estimators defined above. In Assumption 6.2(b), the "0" can be replaced with "$o_p(\sqrt{n})$" without affecting any of the results below.

Now, we are ready to state the theorem for the moment-based models, which says that all the results in the previous sections carry over to the moment-based models:

Theorem 6.1. *With Assumption* 3.1 *replaced by* 6.1, *with* $\mathcal{P}$ *and* $\mathcal{P}_0$ *defined in Definitions* 6.1-6.2, *with Assumption* 6.2 *added, and with all symbols involved taking their new meanings acquired in this section, Theorems* 3.1-5.1 *remain valid.*

[10] The consistency results in these two papers are under fixed DGPs. Minor extensions to allow drifting sequences of DGPs can be made similarly to Shi (2009).

Proof. The proofs of the new theorem are the same as those for the old ones except when Lemma A.1(a) and when the first-order conditions of the likelihood maximization problem are used. In the former case, use Assumption 6.2(a) instead. In the latter case, use condition (i) of Definition 6.1 and Assumption 6.2(b) instead. □

7 Simulation Examples

In this section, I illustrate the performance of the classical Vuong tests and the new nondegenerate test. Three examples are considered. The first example is a prototype of the first source of size distortion: bias in the numerator. The second and third examples are prototypes of the second source of size distortion: the random denominator. The third example is reported in Supplemental Appendix C to save space. In all examples, the number of Monte Carlo repetitions is 5000 and the number of random draws used to obtain simulated critical values is 5001.

The main observations from the examples are (i) both the classical one-step and two-step Vuong tests can generate serious size distortions at all sample sizes, and (ii) the new nondegenerate test has little size distortion. The first example also shows that the nondegenerate test has a substantial power advantage when the more parsimonious competing model is the better one and it has good power when the opposite is true as well. The second example also illustrates the trade-off of size and power between the one-step and the two-step tests and that the nondegenerate test is not subject to the same trade-off. The third example neatly demonstrates the finite sample implications of the theoretical local power results.

For comparison, I also report simulation results for Schennach and Wilhelm's (2011) split-sample test (SW test for short).[11] The results show that the SW test with the robust weighting recommended in the 2011 version of their paper has less power than my test most of the time, significantly less some of the time. It is also over-sized in several scenarios.

To implement the nondegenerate test, I closely follow the steps in Section 4.2. To implement Schennach and Wilhelm's (2011) test, I use their robust weighting parameter.

7.1 Normal Regression Example

Now consider a setup similar to Example 1 introduced in Section 3:

$$\mathcal{F}: Y = \theta^{(0)} + \sum_{j=1}^{K_f} \theta^{(j)} Z_{f,j} + v,\ v|\vec{Z}_f, \vec{Z}_g \sim N(0, \sigma_f^2);$$
$$\mathcal{G}: Y = \beta^{(0)} + \sum_{j=1}^{K_g} \beta^{(j)} Z_{g,j} + u,\ u|\vec{Z}_f, \vec{Z}_g \sim N(0, \sigma_g^2).$$

[11] Because the paper by Schennach and Wilhelm (2011) is not publicly available at this point, a brief description of their test is given in Supplemental Appendix D. I thank them for sending their manuscript and checking my description.

The models now do not assume known error variance. Consider DGPs of the form:

$$Y = 1 + \frac{a_1}{\sqrt{K_f}}\left(\sum_{j=1}^{K_f} Z_{f,j}\right) + \frac{a_2}{\sqrt{K_g}}\left(\sum_{j=1}^{K_g} Z_{g,j}\right) + \varepsilon$$
$$\left(Z_{f,1}, \cdots, Z_{f,K_f}, Z_{g,1}, \cdots, Z_{g,K_g}, \varepsilon\right) \sim N\left(0, I_{K_f+K_g+1}\right) \tag{7.1}$$

Under this DGP, $E\left[\Lambda_i\left(\phi^*\right)\right] = (\log(1+a_1^2) - \log(1+a_2^2))/2$ and ω^2 is determined by the magnitude of both a_1 and a_2.

I consider a base case with $n = 250$, $K_f = 1$ and $K_g = 9$.[12] I also consider four variants of the base case. Each variant differs from the base case in one and only one aspect. The first and the second variants have a larger (19) and a smaller (4) K_g respectively. The third and the fourth variants use a larger (500) and a smaller (100) sample size respectively.

For all variants and the base case, I consider a null and two local alternative DGPs. The null DGP is obtained by setting $a_1 = a_2 = 0.25$. The first alternative DGP sets $a_1 = \sqrt{1.09^{250/n} - 1}$, $a_2 = 0$. Under this DGP, model $\mathcal{F}$ is the true model, model $\mathcal{G}$ is not, and $E_{P_n}[\Lambda_i(\phi^*)] = \log(1.09) \times (250/n)$. The second alternative DGP is the opposite and sets $a_1 = 0$, $a_2 = \sqrt{1.09^{250/n} - 1}$, and under this DGP, $E_{P_n}[\Lambda_i(\phi^*)] = -\log(1.09) \times (250/n)$.

Table 1 shows the rejection probabilities of the classical Vuong tests (1-Step T, 2-Step T), my new nondegenerate test (ND test) and the SW test. The nominal size of all tests is 5%. The rejection probabilities are reported as pairs (p_1, p_2), where p_1 is the probability of rejecting H_0 and choosing $\mathcal{F}$ and p_2 is that of choosing $\mathcal{G}$. The sum of p_1 and p_2 is the rejection probability. In Table 1, I also report the rejection probabilities of the variance test (that is, the first step of the two-step classical Vuong test, labelled "Var.T" in the table).

The first panel of Table 1 shows the null rejection probabilities. As can be seen from this panel, the one-step Vuong test over-rejects (8% vs. 5%) in the base case. The over-rejection becomes more severe as K_g increases to 19 (28% vs. 5%) as predicted by Theorem 3.2. Over-rejection is evident in the variant with ($n = 100$) as well. The two-step Vuong test and the SW test have less over-rejection in all cases, though their over-rejection is apparent in some cases. On the contrary, my nondegenerate test does not over-reject in any of the cases considered.

The second panel of Table 1 shows the rejection probabilities when the more parsimonious model $\mathcal{F}$ is the true model and model $\mathcal{G}$ is wrong. The probabilities that the classical Vuong tests or the SW test picks the correct model are lower than those for the nondegenerate test, and a lot lower in all the cases except when the dimensions of the two models are close ($K_g = 6$).

The third panel shows rejection probabilities when the less parsimonious model $\mathcal{G}$ is the true model and model $\mathcal{F}$ is wrong. Not surprisingly, the classical Vuong tests pick the correct model with high probabilities, but as we see from the first panel, this is because these tests tend to always pick the less parsimonious model, even when they should not. My nondegenerate test has rejection rates comparable to the second panel, showing that it has symmetric power under

[12] K_f and K_g are set so that model $\mathcal{F}$ and model $\mathcal{G}$ have 2 and 10 regressors, respectively.

Table 1: Rejection Probabilities of Different Tests for the Normal Regression Example

$\frac{n}{250}E\Lambda_i(\phi^*)$	Cases	1-Step T	2-Step T	Var.T	ND Test	SW Test
0	Base	$(.00, .08)^*$	$(.00, .08)$	.95	$(.02, .01)$	$(.01, .06)$
	$K_g = 19$	$(.00, .28)$	$(.00, .19)$	.67	$(.01, .01)$	$(.00, .19)$
	$K_g = 4$	$(.01, .04)$	$(.01, .04)$	.99	$(.01, 01)$	$(.02, .04)$
	$n = 100$	$(.00, .13)$	$(.00, .05)$	.27	$(.01, .01)$	$(.00, .12)$
	$n = 500$	$(.00, .07)$	$(.00, .07)$	1.00	$(.02, .02)$	$(.01, .05)$
$\log 1.09$	Base	$(.17, .00)$	$(.16, .00)$	.82	$(.43, .00)$	$(.12, .00)$
	$K_g = 19$	$(.02, .00)$	$(.02, .00)$	.44	$(.34, .00)$	$(.03, .01)$
	$K_g = 4$	$(.44, .00)$	$(.44, .00)$	.95	$(.51, .00)$	$(.22, .00)$
	$n = 100$	$(.22, .00)$	$(.16, .00)$	.58	$(.42, .00)$	$(.20, .00)$
	$n = 500$	$(.17, .00)$	$(.17, .00)$	.87	$(.45, .00)$	$(.09, .01)$
$-\log 1.09$	Base	$(.00, .90)$	$(.00, .79)$	.82	$(.00, .40)$	$(.00, .45)$
	$K_g = 19$	$(.00, .99)$	$(.00, .43)$	.43	$(.00, .34)$	$(.00, .70)$
	$K_g = 4$	$(.00, .79)$	$(.00, .79)$	.95	$(.00, .48)$	$(.00, .34)$
	$n = 100$	$(.00, .93)$	$(.00, .57)$	.58	$(.00, .44)$	$(.00, .78)$
	$n = 500$	$(.00, .90)$	$(.00, .83)$	.87	$(.00, .39)$	$(.00, .27)$

*: (p_1,p_2) = (Pr(rejecting H_0 and selecting $\mathcal{F}$), Pr(rejecting H_0 and selecting $\mathcal{G}$)).

symmetric deviations from the null in both directions. In other words, my test does not bias against either model.

One last thing worth noting is that the power of the nondegenerate Vuong test stays roughly constant as we increase the sample size from 100 to 500 while keeping nLR_{P_0} constant. This is consistent with the prediction of the n^{-1} local power result, Theorem 5.1(b). On the other hand, the power of the SW test decreases noticeably as n increases.

7.2 Joint Normal Location Example

Next, we consider an example designed to illustrate the size distortion effect of the randomness in the denominator in the classical Vuong-test statistic. As the numerical example in Section 3.3 shows, such a size distortion effect is apparent in the $\rho^* = (1, 0)$ case. The example thus is designed to generate a ρ^* close to $(1, 0)$, that is, a case where the log density ratio is highly correlated with the $\partial \log f_i(\theta^*_{P_0})/\partial\theta$, but nearly uncorrelated with $\partial \log g_i(\beta^*_{P_0})/\partial\beta$. To make the task easy, I consider a simple data structure, where there is a two-dimensional observable: $Y = (Y_1, Y_2)'$, and each model contains only one parameter.

Example 2 (Joint Normal Location)**.**

$$\mathcal{F} : (Y_1, Y_2) \sim N\left((\theta, 0), I_2\right), \theta \in R;$$
$$\mathcal{G} : (Y_1, Y_2) \sim N\left((0, \beta), I_2\right), \beta \in R.$$

To generate the data, I let $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 \\ \beta_0 \end{pmatrix}, \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}\right)$. It is easy to calculate that the log-likelihood ratio of the two models under such a DGP is $E[\Lambda_i(\phi^*)] = \left(\theta_0^2 - \beta_0^2\right)$ and the variance of the log-density ratio $\omega^2 = 25\theta_0^2 + \beta_0^2$. Thus, the null hypothesis H_0 holds if and only if $\theta_0^2 = \beta_0^2$, and the asymptotic variance ω^2 increases with both θ_0^2 and β_0^2. Under H_0, calculation shows that $\rho^* = (5/\sqrt{26}, -1/\sqrt{26})' \approx (0.98, -0.20)'$, which is relatively close to the extreme value $(1, 0)'$.

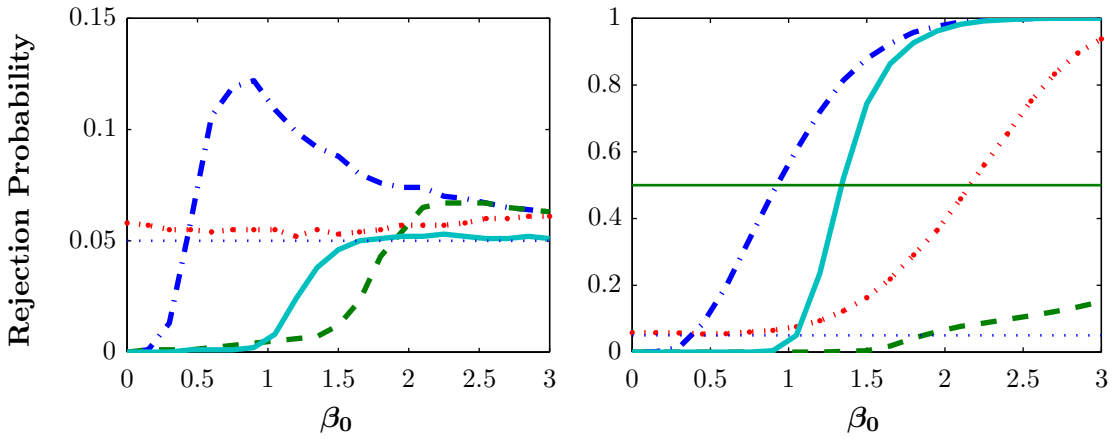
To gain a complete picture of the size of the different tests (of nominal size 5%), I plot the rejection probabilities under DGPs with $\theta_0 = \beta_0$ and $\beta_0 \in [0, 3]$. These rejection probabilities for four different tests and three sample sizes are plotted against β_0 in the three subplots on the left of Figure 3 on the next page. It is easy to see that both my new nondenerate test (solid line) and the SW test (dashed line) achieve good size control – their rejection probabilities stay close to or below the nominal size 5%. On the other hand, the one-step Vuong test (dash-dotted line) has large over-rejection (about 12% as opposed to the nominal size 5%) at all three sample sizes considered. The two-step Vuong test has much smaller, but still some, over-rejection. Increasing the sample size does not reduce the maximum over-rejection of either the one-step Vuong test or the two-step Vuong test, but only changes where the maximum over-rejection occurs.

To compare the power of different tests, I plot the rejection probabilities under DGPs with $\theta_0 = 0$ and $\beta_0 \in [0, 3]$. As we can see, the two tests that have good size properties – my nondegenerate test and the SW test – do not have the same power property. My test outperforms the SW test for all the β_0 values at which both tests have nontrivial power. For example, to achieve a 50% rejection rate, β_0 needs to be about 0.7 for the nondegenerate test, while it needs to be twice as large for the SW test. Among the four tests, the two-step Vuong test has by far the poorest power.

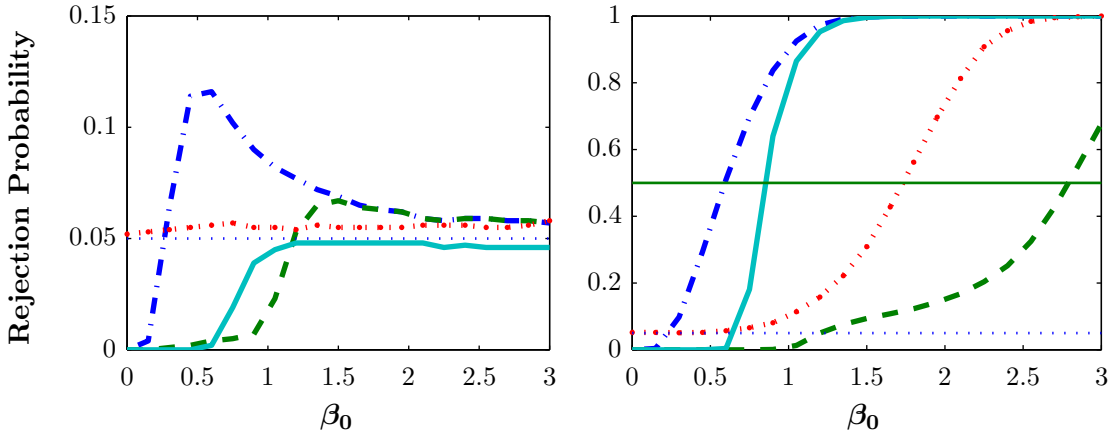
8 Empirical Application to Voter Turnout

In this section, I apply my new test to models of voter turnout using the Texas liquor referenda data collected by Stephen Coate and Michael Conlin. The exercise aims to illustrate the implementation of the new test in STATA and to investigate the extent to which the new test yields conclusions different from those of the classical Vuong tests.

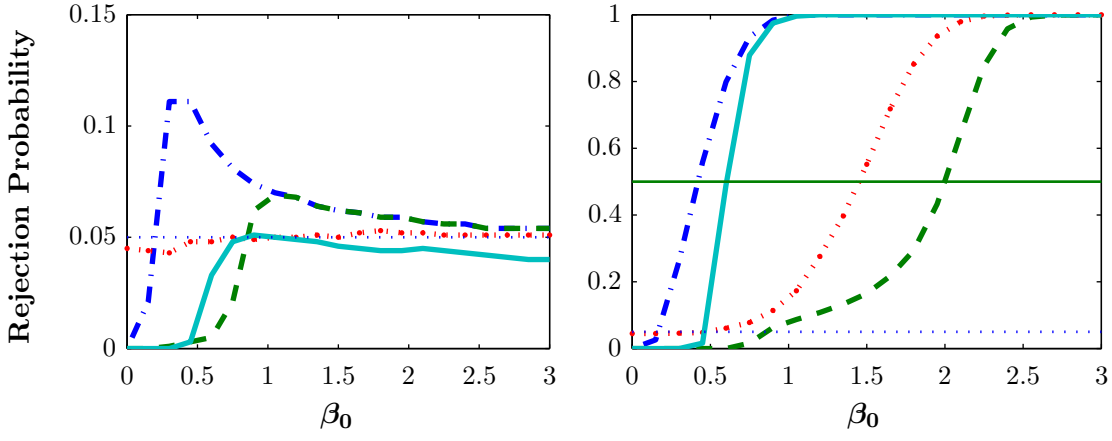
The question studied in Coate and Conlin (2004) is how people decide to vote or not in an election. This is a central problem in political economy. Many theoretical models can be built. When multiple theories are plausible, the Vuong test is a convenient tool to evaluate them in real data. Coate and Conlin (2004) use the classical one-step Vuong test to evaluate the relative fit of three models on the Texas liquor referenda data. The three models evaluated are the linear probability reduced-form model, the intensity model and the group-rule utilitarian model. They find that the group-rule utilitarian model performs significantly better. That is, when testing the null hypothesis that the group-rule utilitarian model and another model are equally distant



(a) $\theta_0 = \beta_0$, $n = 100$ (b) $\theta_0 \neq \beta_0$, $n = 100$



(c) $\theta_0 = \beta_0$, $n = 250$ (d) $\theta_0 \neq \beta_0$, $n = 250$



(e) $\theta_0 = \beta_0$, $n = 500$ (f) $\theta_0 \neq \beta_0$, $n = 500$

Figure 3: Rejection probabilities of the one-step Vuong test (dash-dotted line), the two-step Vuong test (dashed line), the SW test (dotted line), and my new nondegenerate test (solid line) for Example 2. The horizontal dotted line indicates the nominal size 5%.

to the true data distribution, the classical Vuong test of nominal level 5% rejects in favor of the former.

First, I briefly describe the Texas liquor referenda data. The data are collected from 363 local liquor elections in Texas between 1976 and 1996. The elections are on proposals that, if passed, would have relaxed liquor regulation in the jurisdiction. Thus, supporting the proposals represents a preference for less liquor regulation. The data set contains the number of support votes, the number of opposing votes, the number of eligible voters in the jurisdiction at election time, demographic and economic information on the jurisdiction as well as weather and other features of the election day. More details and summary statistics are in Coate and Conlin (2004).

I consider all three models studied in Coate and Conlin (2004). To make this section self-contained, I describe these models briefly here. Greater detail and background information can be found in Coate and Conlin (2004). The *reduced-form* model (RF for short) is a seemingly unrelated regression model in which the fraction of the eligible voters who voted to support the proposal and that of the eligible voters who voted to oppose it are dependent variables and all covariates mentioned above are explanatory variables.

In the *group rule-utilitarian* model (GRU for short), the eligible voters live on a continuum and incur a heterogeneous cost for voting. They are exogenously divided into two groups: supporters and opposers. Each supporter derives a positive utility b from having the proposal passed, while each opposer derives a positive utility x from having the proposal blocked. Each group (supporters and opposers) collectively and simultaneously chooses a rule (that dictates which members vote and which stay home) to maximize the aggregate utility of its members given the rule of the other group. The voting behavior will be a result of an equilibrium of the game between the two groups. The "group" dictator here serves as an approximation to people's rule-utilitarian behavior – individuals follow the rule that would maximize the aggregate expected utility of the group they belong to if everybody in the group followed it. The election is won by the side that sends more people to vote. Coate and Conlin (2004) estimated the model after parameterizing the winning benefits, the voting cost, and the distribution of supports in the population of eligible voters.

In the *intensity* model (INT for short) all model primitives are assumed to be the same as those of the GRU model. But instead of being rule-utilitarian, each individual supporter simply chooses to vote if her voting cost is below αb and each individual opposer chooses to vote if his voting cost is below αx, where α is the intensity with which the individuals care about the election outcome.

I first estimate the models in STATA using its official "ml" package and obtain the covariance of the scores (matrix $\hat{B}_n$) as well as the Hessian (matrix $\hat{A}_n$) in the following way.[13] The scores are conveniently available from the "ml score" command, while the inverse of the Hessian is

[13]For estimation, I use the STATA code of Coate and Conlin (2004), downloaded from the website of the American Economic Review. The cleaned data set was from the same source. We were able to use their code without modification and to replicate the tables in their paper with remarkable precision.

simply the variance-covariance matrix of the parameter estimators reported in the STATA macro "e(V)" after "ml" estimation. Using this set of information, I then compute the test statistic and the critical value of the new test following the procedures described in Section 4.2 using Mata.

I apply the classical Vuong test and the new nondegenerate Vuong test to the pairwise comparison of the three models described above. The results are reported in Table 2. Because the nondegenerate test differs from the classical Vuong test both in test statistic and in critical value, for a more informative comparison, I report the p-values for both tests.[14] The result of the nondegenerate test confirms Coate and Conlin's (2004) conclusion that the GRU model is sigifnicantly closer to the truth than the RF model at levels lower than 1%. However, the evidence for the other model comparison becomes weaker. While Coate and Conlin (2004) find that the GRU model is significantly closer to the truth than the INT model using the classical Vuong test, the conclusion does not hold even at the 10% level according to the new test.

Table 2: Results of the Nondegenerate Vuong Test and the Classical Vuong Test

$\mathcal{F}$(Log-likelihood)	$\mathcal{G}$(Log-likelihood)	p-value of Nondegenerate Test	p-value of Classical Vuong Test
GRU(748.59)	INT(706.41)	.142	.037**
GRU(748.59)	NF(662.90)	.003***	.001***
INT(706.41)	NF(662.90)	.073*	.105

Note: "GRU" stands for "group rule-utlitarian" model, "INT" stands for the "Intensity" model and "RF' stands for the "reduced-form" model. The tests are for $H_0 : LL(\mathcal{F}) = LL(\mathcal{G})$ against $H_1 : LL(\mathcal{F}) \neq LL(\mathcal{G})$. The "***", "**", and "*" indicate statistical significance at the 1%, 5% and 10% levels, respectively.

It is worth noting that the nondegenerate Vuong test does not always yield weaker significance – in the comparison between the intensity model and the linear probability reduced model, the p-value of the nondegenerate Vuong test is in fact smaller than that of the classical Vuong test. This suggests that the kind of difference that my proposed bias correction and the variance adjustment make depends on the empirical context.

9 Conclusion

To sum up, this paper proposes a new nondegenerate test as an alternative and a modification of the classical Vuong tests, and extends it to moment-based models. The analysis complements Vuong (1989) as well as Kitamura (2000) by making their tests rigorous in a uniform sense, and opens the door for further research on uniform inference in nonnested testing problems. The test is implemented on a voter turnout data set from Coate and Conlin (2004).

[14] To find the p-value of the nondegenerate test, I try different significance levels and find the lowest one for which the test statistic still exceeds the critical value. The p-values reported are accurate to the third digit.

Appendix

A Notation and Auxiliary Lemmas

Let "LLN" denote the weak law of large numbers for row-wise i.i.d. triangular arrays. I use Theorem 2 in Andrews (1988). This theorem is a law of large numbers for L^1-mixingale triangular arrays, which includes row-wise i.i.d. triangular arrays as a special case. The uniform integrability condition in that theorem is guaranteed by moment conditions in this paper. Let "ULLN" denote the uniform weak law of large numbers. I use Theorem 4 in Andrews (1992). This theorem covers the case of i.n.i.d. sequences instead of row-wise i.i.d. triangular arrays of random vectors, but all relevant proofs in that paper go through for the latter. Let "CLT" denote the Lyapounov central limit theorem. Let $\Lambda_{i,n}^* = \Lambda_i(\phi_{P_n}^*)$. Let "wp$\to$ 1" denote "with probability approaching one." Let "f.o.c." stand for "first-order condition."

Lemma A.1 below shows the consistency of the estimators of matrices A and B. Lemma A.2 gives the asymptotics of the main components in $\widehat{LR}_n$ and $\hat{\omega}_n^2$ and is used in the proof of Theorem 3.1.

Lemma A.1. *Suppose Assumption* 3.1 *holds, under a sequence* $\{P_n\}_{n=1}^{\infty} \in Seq(\sigma^2, A, B, \rho)$ *for some* (σ^2, A, B, ρ)*, if the random sequence* $\tilde{\phi}_n$ *satisfies* $\|\tilde{\phi}_n - \phi_{P_n}^*\| \to_p 0$*, then we have*

$$\left(n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \Lambda_i(\tilde{\phi}_n)}{\partial \phi \partial \phi'}, n^{-1} \sum_{i=1}^{n} \frac{\partial \Lambda_i(\tilde{\phi}_n)}{\partial \phi} \frac{\partial \Lambda_i(\tilde{\phi}_n)}{\partial \phi'} \right) \to_p (A, B).$$

Proof of Lemma A.1. I focus on $n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \Lambda_i(\tilde{\phi}_n)}{\partial \phi \partial \phi'}$ because the other part of the lemma follows from similar arguments.

First, by the ULLN, we have

$$\sup_{\phi \in \Theta \times \mathcal{B}} \left\| n^{-1} \sum\nolimits_{i=1}^{n} \partial^2 \Lambda_i(\phi) / \partial \phi \partial \phi' - A_{P_n}(\phi) \right\| \to_p 0, \tag{A.1}$$

where the ULLN applies because the four sufficient conditions in it are satisfied. In particular, the total boundedness condition (BD) in the ULLN holds by Assumption 3.1(b), the pointwise convergence condition (P-WLLN) is guaranteed by the LLN, the domination condition (DM) is guaranteed by condition (iii) of Definiton 3.1, and the termwise stochastic equicontinuity (TSE) condition holds because

$$\begin{aligned} &E_{P_n} \sup\nolimits_{\phi, \phi' \in \Theta \times \mathcal{B}: \|\phi - \phi'\| < d} \left\| \partial^2 \Lambda_i(\phi) / \partial \phi \partial \phi' - \partial^2 \Lambda_i(\phi') / \partial \phi \partial \phi' \right\| \\ &\leq E_{P_n} \sup\nolimits_{\phi \in \Theta \times \mathcal{B}} \sum\nolimits_{j=1}^{d_\theta + d_\beta} \left\| \partial^3 \Lambda_i(\phi) / \partial \phi_j \partial \phi \partial \phi' \right\| \cdot d \leq M \cdot d, \end{aligned} \tag{A.2}$$

where the first inequality holds by a mean-value expansion and the second holds by condition (iii) of Definition 3.1. (See Andrews (1992) (Thm. 4) for the details of the ULLN conditions.)

In addition to guaranteeing TSE, equation (A.2) also shows the uniform continuity of $\partial\Lambda_i(\phi)/\partial\phi\partial\phi'$ in $\Theta \times \mathcal{B}$. This, $\|\tilde{\phi}_n - \phi_{P_n}\| \to_p 0$, equation (A.1) and Definition 3.3 together show the desired result. □

Lemma A.2. *Suppose Assumption* 3.1 *holds. Under a drifting sequence* $\{P_n\}_{n=1}^{\infty} \in Seq(\sigma^2, A, B, \rho)$ *for some* (σ^2, A, B, ρ)*, then* (a) $\|\hat{\phi}_n - \phi^*_{P_n}\| \to_p 0$, (b) $n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}) \to_d N(0, A^{-1}BA^{-1})$, (c) *if* $\sigma \in [0, \infty)$,

$$\begin{pmatrix} \sum_{i=1}^n (\Lambda^*_{i,n} - E_{P_n}\Lambda^*_{i,n}) \\ n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}) \end{pmatrix} \to_d \begin{pmatrix} \sigma Z_\Lambda \\ -A^{-1}Z_\phi \end{pmatrix}, \text{ and}$$

(d) *if* $\sigma = \infty$, $n^{-1/2}\omega_{P_n}^{-1}\sum_{i=1}^n(\Lambda^*_{i,n} - E_{P_n}\Lambda^*_{i,n}) \to_d N(0,1)$.

Proof of Lemma A.2. (a) I only show the consistency of $\hat{\theta}_n$ because that of $\hat{\beta}_n$ follows by analogous arguments. Let $ll_{f,n}$ stand for ll_{f,P_n}. Let the sample log-likelihood function be denoted

$$\widehat{ll}_{f,n}(\theta) = n^{-1}\sum_{i=1}^n \log f_i(\theta). \tag{A.3}$$

The consistency of $\hat{\theta}_n$ is implied by the following derivation. For any $\varepsilon > 0$ and $\delta(\varepsilon)$ in Definition 3.1,

$$\begin{aligned}
&\Pr_{P_n}\left(\|\hat{\theta}_n - \theta^*_{P_n}\| > \varepsilon\right) \\
\leq &\Pr_{P_n}\left(ll_{f,n}(\theta^*_{P_n}) - ll_{f,n}(\hat{\theta}_n) > \delta(\varepsilon)\right) \\
= &\Pr_{P_n}\left(ll_{f,n}(\theta^*_{P_n}) - \widehat{ll}_{f,n}(\theta^*_{P_n}) + \widehat{ll}_{f,n}(\theta^*_{P_n}) - \widehat{ll}_{f,n}(\hat{\theta}_n) + \widehat{ll}_{f,n}(\hat{\theta}_n) - ll_{f,n}(\hat{\theta}_n) > \delta(\varepsilon)\right) \\
\leq &\Pr_{P_n}\left(\sup_{\theta\in\Theta}|ll_{f,n}(\theta) - \widehat{ll}_{f,n}(\theta)| > \delta(\varepsilon)/2\right) \to 0,
\end{aligned} \tag{A.4}$$

where the first inequality holds by condition (i) of Definition 3.1, the second inequality holds because $\hat{\theta}_n$ maximizes $\widehat{ll}_{f,n}(\theta)$, and the convergence holds because $\sup_{\theta\in\Theta}|ll_{f,n}(\theta) - \widehat{ll}_{f,n}(\theta)| \to_p 0$, which holds by arguments similar to those for (A.1) in the proof of Lemma A.1.

(b) By mean-value expansions of the f.o.c.s from log-likelihood maximization,

$$\begin{aligned}
0 &= n^{-1/2}\sum\nolimits_{i=1}^n \partial\Lambda_i(\phi^*_{P_n})/\partial\phi + \left(\partial^2\Lambda_i(\tilde{\phi}_n)/\partial\phi\partial\phi'\right) n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}). \\
&= n^{-1/2}\sum\nolimits_{i=1}^n \partial\Lambda_i(\phi^*_{P_n})/\partial\phi + (A + o_p(1))n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}),
\end{aligned} \tag{A.5}$$

where $\tilde{\phi}_n$ is a point lying on the line segment joining $\hat{\phi}_n$ and $\phi^*_{P_n}$, and the second equality holds by Lemma A.1 and Lemma A.2(a).

By conditions (i)-(ii) of Definition 3.1, we have $E_{P_n}\partial\Lambda_i(\phi^*_{P_n})/\partial\phi = 0$. Then, we can appeal to the CLT and get

$$n^{-1/2}\sum\nolimits_{i=1}^n \partial\Lambda_i(\phi^*_{P_n})/\partial\phi \to_d N(0, B). \tag{A.6}$$

Equations (A.5), (A.6), condition (v) of Definition 3.1 and the Δ-method combined produce the desired result.

(c) I treat the two cases: (i) $\sigma = 0$ and (ii) $\sigma \in (0, \infty)$ separately. (i) $\sigma = 0$. Because $E_n(\sum_{i=1}^n (\Lambda_{i,n}^* - E_{P_n}\Lambda_{i,n}^*))^2 = n\omega_{P_n}^2 \to 0$, we have $\sum_{i=1}^n (\Lambda_{i,n}^* - E_{P_n}\Lambda_{i,n}^*) \to_p 0$. This combined with part (b) shows part (c).

(ii) $\sigma \in (0, \infty)$. We have

$$E_{P_n}\begin{pmatrix} n^{1/2}(\Lambda_{i,n}^* - E_{P_n}\Lambda_{i,n}^*) \\ \partial\Lambda_i(\phi_{P_n}^*)/\partial\phi \end{pmatrix} = 0,$$
$$Var_{P_n}\begin{pmatrix} n^{1/2}\Lambda_{i,n}^* \\ \partial\Lambda_i(\phi_{P_n}^*)/\partial\phi \end{pmatrix} \to \begin{pmatrix} \sigma^2 & \sigma\rho' D^{1/2} \\ \sigma D^{1/2}\rho & B \end{pmatrix}, \tag{A.7}$$

where the first equality holds by $P_n \in \mathcal{P}_0$ and by the first-order conditions from maximizing the population log-likelihood, and the convergence holds by Definition 3.3.

Because $\sigma > 0$, we have $\omega_{P_n}^2 > 0$ for all large enough n. Given this, conditions (iii) and (iv) of Definition 3.1 ensure that the Lyapounov condition holds. Thus, we can apply the CLT and obtain

$$\begin{pmatrix} \sum_{i=1}^n (\Lambda_{i,n}^* - E_{P_n}\Lambda_{i,n}^*) \\ n^{-1/2}\sum_{i=1}^n \partial\Lambda_i(\phi_{P_n}^*)/\partial\phi \end{pmatrix} \to_d \begin{pmatrix} \sigma Z_\Lambda \\ Z_\phi \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & \sigma\rho' D^{1/2} \\ \sigma D^{1/2}\rho & B \end{pmatrix}\right). \tag{A.8}$$

Then, (A.5), (A.8), condition (v) of Definition 3.1 and the Δ-method conclude the proof.

(d) The proof is omitted because it is similar to and simpler than part (c). □

B Proof of Main Results

Proof of Theorem 3.1. (a) First, I derive the asymptotic distribution of $n\widehat{LR}_n$. A Taylor expansion of $\Lambda_i(\phi_{P_n}^*)$ around $\hat{\phi}_n$ gives

$$\begin{aligned} n\widehat{LR}_n &= \sum_{i=1}^n \left[\Lambda_{i,n}^* - \frac{\partial\Lambda_i(\hat{\phi}_n)}{\partial\phi'}(\hat{\phi}_n - \phi_{P_n}^*)\right] - n(\hat{\phi}_n - \phi_{P_n}^*)' \left[n^{-1}\sum_{i=1}^n \frac{\partial^2\Lambda_i(\tilde{\phi}_n)}{\partial\phi\partial\phi'}\right](\hat{\phi}_n - \phi_{P_n}^*)/2 \\ &= \sum_{i=1}^n \Lambda_{i,n}^* - 2^{-1}n^{1/2}(\hat{\phi}_n - \phi_{P_n}^*)' \left[n^{-1}\sum_{i=1}^n \frac{\partial^2\Lambda_i(\tilde{\phi}_n)}{\partial\phi\partial\phi'}\right] n^{1/2}(\hat{\phi}_n - \phi_{P_n}^*) \\ &\equiv n\widehat{LR1}_n + n\widehat{LR2}_n, \end{aligned} \tag{B.1}$$

where $\tilde{\phi}_n$ lies on the line segment joining $\phi_{P_n}^*$ and $\hat{\phi}_n$, and the second equality holds wp$\to$ 1 by the f.o.c. from the likelihood maximization problem and condition (ii) of Definition 3.1. The joint asymptotic distribution of $n\widehat{LR1}_n$ and $n^{1/2}(\hat{\phi}_n - \phi_{P_n}^*)$ is given in Lemma A.2(c). Lemma

A.1 and Lemma A.2(a) together show that

$$n^{-1}\sum_{i=1}^{n}\partial^2\Lambda_i(\tilde{\phi}_n)/\partial\phi\partial\phi' \to_p A. \tag{B.2}$$

Combining (B.2) and Lemma A.2(c), we have $n\widehat{LR}_n \to_d \sigma^{1/2}Z_\Lambda - 2^{-1}Z_\phi' A^{-1}Z_\phi$.

Now, I derive the asymptotic distribution of $n\hat{\omega}_n^2$. A mean-value expansion of $\Lambda_i(\hat{\phi}_n)$ around $\phi_{P_n}^*$ gives, for some $\tilde{\phi}_n$ lying on the line segment joining $\phi_{P_n}^*$ and $\hat{\phi}_n$,

$$\begin{aligned} n\hat{\omega}_n^2 =& -n\widehat{LR}_n^2 + \sum_{i=1}^{n}(\Lambda_{i,n}^*)^2 + 2\sum_{i=1}^{n}\Lambda_{i,n}^*\cdot(\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi')(\hat{\phi}_n-\phi_{P_n}^*) \\ &+ (\hat{\phi}_n-\phi_{P_n}^*)'\sum_{i=1}^{n}(\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi)(\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi')(\hat{\phi}_n-\phi_{P_n}^*) \\ =& o_p(1) + W_{n,1} + 2W_{n,2}\cdot n^{1/2}(\hat{\phi}_n-\phi_{P_n}^*) + n^{1/2}(\hat{\phi}_n-\phi_{P_n}^*)'W_{n,3}\cdot n^{1/2}(\hat{\phi}_n-\phi_{P_n}^*), \end{aligned} \tag{B.3}$$

where the $o_p(1)$ comes from $n\widehat{LR}_n = O_p(1)$ shown above. For the W terms, I treat the two cases (i) $\sigma = 0$ and (ii) $\sigma \in (0,\infty)$ separately below.

(i) $\sigma = 0$. First, because $E_{P_n}|W_{n,1}| = E_{P_n}W_{n,1} = n\omega_{P_n}^2 \to 0$, we have

$$W_{n,1} \equiv \sum_{i=1}^{n}(\Lambda_{i,n}^*)^2 = o_p(1) \tag{B.4}$$

Second, by the Cauchy-Schwarz inequality, we have

$$\|W_{n,2}\| \le W_{n,1}^{1/2}\cdot\left[n^{-1}\sum_{i=1}^{n}\|\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi'\|^2\right]^{1/2} = o_p(1)\cdot O_p(1) = o_p(1), \tag{B.5}$$

where the first equality holds by (B.4) and the inequality $E_{P_n}n^{-1}\sum_{i=1}^{n}\|\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi'\|^2 \le E_n\sup_\phi\|\partial\Lambda_i(\phi)/\partial\phi'\|^2 < 2M$, which holds by condition (iii) of Definition 3.1. Third,

$$W_{n,3} \equiv n^{-1}\sum_{i=1}^{n}\frac{\partial\Lambda_i(\tilde{\phi}_n)}{\partial\phi}\frac{\partial\Lambda_i(\tilde{\phi}_n)}{\partial\phi'} \to_p B, \tag{B.6}$$

by Lemmas A.1 and A.2(a). By (B.3)-(B.6) and Lemma A.2(b), we have

$$n\hat{\omega}_n^2 \to_d Z_\phi' A^{-1}BA^{-1}Z_\phi. \tag{B.7}$$

(ii) $\sigma \in (0,\infty)$. Because $\sigma > 0$, $\omega_{P_n}^2 > 0$ for n large enough. Thus, the triangular array $\{(\omega_{P_n}^{-1}\Lambda_{i,n}^*)^2\}_{i\le n,\, n\ge 1}$ is uniformly integrable by condition (iv) of Definition 3.1. With the uniform

integrability and $E(\omega_n^{-1}\Lambda_{i,n}^*)^2 = 1$, the LLN applies and gives:

$$n^{-1}\omega_{P_n}^{-2}W_{n,1} = n^{-1}\sum_{i=1}^{n}(\omega_{P_n}^{-1}\Lambda_{i,n}^*)^2 \to_p 1. \tag{B.8}$$

By similar arguments,

$$\begin{aligned} n^{-1}\omega_{P_n}^{-2}W_{n,2} &\equiv n^{-1}\omega_{P_n}^{-2}\cdot n^{-1/2}\sum_{i=1}^{n}\Lambda_{i,n}^*\cdot\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi' \\ &= n^{-1/2}\omega_{P_n}^{-1}\cdot n^{-1}\sum_{i=1}^{n}\omega_{P_n}^{-1}\Lambda_{i,n}^*\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi' \\ &= (\sigma^{-1}+o(1))n^{-1}\sum_{i=1}^{n}\omega_{P_n}^{-1}\Lambda_{i,n}^*\partial\Lambda_i(\tilde{\phi}_n)/\partial\phi' \\ &= (\sigma^{-1}+o(1))(\rho_n' D_n^{1/2}+o_p(1)) \to_p \sigma^{-1}\rho' D^{1/2}, \end{aligned} \tag{B.9}$$

where the third equality holds by Definition 3.3 and the last equality holds by the same arguments as in the proof of Lemma A.1 except with condition (iii) supplemented with condition (iv) of Definition 3.1 and the Cauchy-Schwarz inequality.

Equation (B.6) remains valid when $\sigma \in (0,\infty)$. Therefore, by (B.3), (B.6), (B.8) and (B.9), we have

$$\omega_{P_n}^{-2}\hat{\omega}_n^2 \to_d 1 - 2\sigma^{-1}\rho' D^{1/2}A^{-1}Z_\phi + \sigma^{-2}Z_\phi' A^{-1}BA^{-1}Z_\phi. \tag{B.10}$$

Equation (B.10) concludes the proof because $n\omega_{P_n}^2 \to \sigma^2 \in (0,\infty)$.

(b) When $\sigma = \infty$, for $\widehat{LR}2_n$ in (B.1), $n^{1/2}\omega_{P_n}^{-1}\widehat{LR}2_n \to_p 0$ because $n\widehat{LR}2_n = O_p(1)$ and $n\omega_{P_n}^2 \to \infty$. For $n\widehat{LR}1_n$ in (B.1), we have

$$n^{1/2}\omega_{P_n}^{-1}\widehat{LR}1_n = n^{-1/2}\sum_{i=1}^{n}\omega_{P_n}^{-1}\Lambda_{i,n}^* \to_d N(0,1), \tag{B.11}$$

by the CLT. The CLT applies by (a) $E_{P_n}\omega_{P_n}^{-1}\Lambda_{i,n}^* = 0$, (b) $E_{P_n}\omega_{P_n}^{-2}(\Lambda_{i,n}^*)^2 = 1$ and also (c) $E_{P_n}\omega_{P_n}^{-2-\delta}|\Lambda_{i,n}^*|^{2+\delta} \le M < \infty$, which holds by condition (iv) of Definition 3.1.

The derivation of the probability limits of $\omega_{P_n}^{-2}\hat{\omega}_n^2$ when $\sigma = \infty$ is the same as when $\sigma \in (0,\infty)$. Simply sending σ to infinity in equation (B.10) gives us the desired result.

(c) First, observe that

$$\begin{aligned} Z_\phi^{*\prime}VZ_\phi^* &= (Z_\phi^{*\prime}Q'B^{1/2})A^{-1}(B^{1/2}QZ_\phi^*) \\ Z_\phi^{*\prime}V^2Z_\phi^* &= Z_\phi^{*\prime}Q'B^{1/2}A^{-1}B^{1/2}QQ'B^{1/2}A^{-1}B^{1/2}QZ_\phi^* \\ &= (Z_\phi^{*\prime}Q'B^{1/2})A^{-1}BA^{-1}(B^{1/2}QZ_\phi^*), \text{ and} \end{aligned}$$

$$\begin{aligned}\rho^{*\prime} V Z_\phi' =& \rho^{*,\prime} Q' B^{1/2} A^{-1} B^{1/2} Q Z_\phi^* \\ =& \rho' D^{1/2} A^{-1} (B^{1/2} Q Z_\phi^*). \end{aligned} \tag{B.12}$$

Let $Z_\phi = B^{1/2} Q Z_\phi^*$. Then the equivalence (in distribution) between the expression in part (c) and that in part (a) becomes apparent. □

Proof of Theorem 3.2. It suffices to show part (a) because part (b) is immediately implied by part (a). For fixed k, we have $Z_\phi^* \hat{V}_n^2 Z_\phi^* \to_d Z_\phi^* V_k^2 Z_\phi^*$ as $n \to \infty$ because $\hat{V}_n \equiv eig(\hat{B}_n^{1/2} \hat{A}_n^{-1} \hat{B}_n^{1/2}) \to_p eig(B_k^{1/2} A_k^{-1} B_k^{1/2}) \equiv V_k$ by Lemma A.1 and Lemma A.2(a). Because $Z_\phi^* V_k^2 Z_\phi^*$ has a continuous and strictly increasing c.d.f., $c_\omega(\hat{V}_n^2, 1-\alpha) \to_p c_\omega(V_k^2, 1-\alpha)$. By this and Theorem 3.1(c), the left-hand side (l.h.s.) of the equation in part (a) equals:

$$\lim_{k\to\infty} \Pr\left(J_{\omega,k} > c_\omega(V_k^2, 1-\alpha) \;\&\; J_{\Lambda,k} \Big/ J_{\omega,k}^{1/2} > z_{\alpha/2} \right), \tag{B.13}$$

where $J_{\Lambda,k} = J_\Lambda(\sigma_k, \rho_k^*, V_k)$ and $J_{\omega,k} = J_\omega(\sigma_k, \rho_k^*, V_k)$. Now we simply need to study the behavior of $J_{\Lambda,k}$, $J_{\Lambda,k} \Big/ J_{\omega,k}^{1/2}$ and $c_\omega(V_k^2, 1-\alpha)$ as $k \to \infty$.

Let t_k stand for $\sqrt{tr(V_k^2)}$. Because $c_\omega(V_k^2, 1-\alpha)$ is the $1-\alpha$ quantile of $Z_\phi^* V_k^2 Z_\phi^*$, $J_\omega > c_\omega(V_k^2, 1-\alpha)$ is equivalent to $J_\omega / t_k^2 > c_\omega(V_k^2/t_k^2, 1-\alpha)$. Consider the derivation, as $k \to \infty$,

$$t_k^{-2} Z_\phi^* V_k^2 Z_\phi^* = t_k^{-2} \sum\nolimits_{j=1}^k v_{k,j}^2 Z_{\phi,j}^{*2} \to_p 1 \tag{B.14}$$

where $v_{k,j}$ is the jth diagonal element of V_k and $Z_{\phi,j}^*$ is the jth element of Z_ϕ^*, and the "$\to_p$" holds because $E\left[t_k^{-2} \sum_{j=1}^k v_{k,j}^2 Z_{\phi,j}^{*2} - 1 \right]^2 = 2tr\left(V_k^4\right) / t_k^4 \to 0$. Thus,

$$c_\omega(V_k^2/t_k^2, 1-\alpha) = O(1), \tag{B.15}$$

Also consider the derivation:

$$\begin{aligned} t_k^{-2} J_{\omega,k} &= t_k^{-2}\sigma_k^2 - 2(\sigma_k/t_k) t_k^{-1} \sum\nolimits_{j=1}^k \rho_j^* v_{k,j} Z_{\phi,j}^* + t_k^{-2} Z_\phi^* V_k^2 Z_\phi^* \\ &= (\sigma_k/t_k)^2 - 2(\sigma_k/t_k) \cdot o_p(1) + 1 + o_p(1) \to_p \infty, \text{ as } k \to \infty. \end{aligned} \tag{B.16}$$

The second $o_p(1)$ in the derivation is obtained by (B.14) and the first $o_p(1)$ is obtained by

$$\begin{aligned} E\left[t_k^{-1} \sum\nolimits_{j=1}^k \rho_j^* v_{k,j} Z_{\phi,j}^* \right]^2 &= t_k^{-2} \sum\nolimits_{j=1}^k \rho_j^{*2} v_{k,j}^2 \le t_k^{-2} \max_j v_{k,j}^2 \\ &\le \left[tr\left(V_k^4\right) / t_k^4 \right]^{1/2} \to 0, \text{ as } k \to \infty, \end{aligned} \tag{B.17}$$

where the first inequality holds because $\|\rho^*\| \le 1$, which holds because $\begin{pmatrix} 1 & \rho^{*\prime} \\ \rho^* & I_{d_\theta + d_\beta} \end{pmatrix}$ is a correla-

tion matrix. Equations (B.15) and (B.16) imply that the expression in (B.13) equals:

$$\lim_{k\to\infty} \Pr\left(J_{\Lambda,k} \Big/ J_{\omega,k}^{1/2} > z_{\alpha/2}\right). \tag{B.18}$$

That is, the pretest rejects wp$\to$ 1. Now observe that

$$\begin{aligned} J_{\Lambda,k} \Big/ J_{\omega,k}^{1/2} &= \frac{Z_\Lambda - (\sigma_k/t_k)^{-1}(2t_k)^{-1}\sum_{j=1}^k v_{k,j}\left(Z_{\phi,j}^{*2}-1\right) - tr\left(V_k\right)/(2\sigma_k)}{\sqrt{1-2\left(\sigma_k/t_k\right)^{-1}t_k^{-1}\sum_{j=1}^k \rho_j^* v_{k,j} Z_{\phi,j}^* + \left(\sigma_k/t_k\right)^{-2} t_k^{-2} Z_\phi^* V_k^2 Z_\phi^*}} \\ &\to_p \infty, \text{ as } k\to\infty \end{aligned} \tag{B.19}$$

where the convergence holds by (B.14), (B.17), $\sigma_k/t_k \to \infty$, $-tr\left(V_k\right)/\left(2\sigma_k\right) \to \infty$, as well as $2^{-1}t_k^{-1}\sum_{j=1}^k v_{k,j}\left(Z_{\phi,j}^{*2}-1\right) \to_d N(0,0.5)$, which holds by the Lyapounov CLT. This implies that the expression in (B.18) equals zero and in turn shows the desired result. □

Proof of Theorem 4.1. We derive the asymptotic size using subsequence arguments similar to those in Andrews, Cheng, and Guggenberger (2009). First, we take a sequence $\{P_n \in \mathcal{P}_0\}$ and a subsequence $\{b_n\}$ of $\{n\}$ such that

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} \Pr{}_P\left(|\hat{T}_n^{\text{mod}}(c)| > cv(1-\alpha,\hat{V}_n,c)\right) \\ &= \lim_{n\to\infty} \Pr{}_{P_{b_n}}\left(|\hat{T}_{b_n}^{\text{mod}}(c)| > cv(1-\alpha,\hat{V}_{b_n},c)\right). \end{aligned} \tag{B.20}$$

Such sequences and subsequences always exist. Condition (iii) of Definition 3.1 implies that elements in the matrices A_P and B_P are uniformly bounded over $P \in \mathcal{P}$. Also, ρ_{P_n}'s elements are between -1 and 1. Thus, there exists a subsequence $\{a_n\}$ of $\{b_n\}$, and some $\left(\sigma^2, A, B, \rho\right)$ such that $\left(a_n\omega_{P_{a_n}}^2, A_{P_{a_n}}, B_{P_{a_n}}, \rho_{P_{a_n}}\right) \to \left(\sigma^2, A, B, \rho\right)$. It suffices to show that

$$\lim_{n\to\infty} \Pr{}_{P_{a_n}}\left(|\hat{T}_{a_n}^{\text{mod}}(c)| > cv(1-\alpha,\hat{V}_{a_n},c)\right) \le \alpha. \tag{B.21}$$

First, I derive the limiting distribution of $\hat{T}_{a_n}^{\text{mod}}(c)$. Lemmas A.1 and A.2(a) imply that $\hat{V}_{a_n} \to_p V \equiv eig(B^{1/2}A^{-1}B^{1/2})$. This combined with Theorem 3.1(c) gives us

$$\hat{T}_{a_n}^{\text{mod}}(c) \to_d J(\sigma,\rho^*,V,c) \equiv \frac{J_\Lambda(\sigma,\rho^*,V) + tr(V)/2}{\sqrt{J_\omega(\sigma,\rho^*,V) + c\cdot tr(V^2)}}, \tag{B.22}$$

where ρ^* is defined in part (c) of Theorem 3.1.

Next, I show that $cv(1-\alpha,V,c)$ is continuous in V. By visual inspection, we see that $J(\sigma,\rho^*,V,c)$ is continuous with probability one in (σ,ρ^*,V) at any point with $\sigma \in [0,\infty]$, $\rho^* \in \{\rho^* \in [0,1]^k : \|\rho^*\| \le 1\}$, and V being a diagonal matrix with real entries. Let $\{V_m\}_{m=1}^\infty$ be a sequence of diagonal matrices that converges to a real diagonal matrix V_∞. Let $\{u_m\}$ be an

arbitrary subsequence of $\{m\}$. Below we show that $\{u_m\}$ has a subsequence $\{a_m\}$ such that

$$cv(1-\alpha, V_{a_m}, c) \to cv(1-\alpha, V_\infty, c) \text{ as } m \to \infty. \tag{B.23}$$

This shows that $cv(1-\alpha, V_m, c) \to cv(1-\alpha, V_\infty, c)$ as $m \to \infty$ because $\{u_m\}$ is an arbitrary subsequence of $\{m\}$. Thus, $cv(1-\alpha, V, c)$ is continuous in V.

Let $\{\sigma_m\}$ and $\{\rho_m^*\}$ be such that $F^{-1}_{|J(\sigma_m, \rho_m^*, V_m, c)|}(1-\alpha) = cv(1-\alpha, V_m, c) + o(1)$ as $m \to \infty$. Such sequences always exist. By the completeness of $[0,\infty]$ and the compactness of $\{\rho^* \in [0,1]^k : \|\rho^*\| \leq 1\}$, there exists a subsequence $\{a_m\}$ of $\{u_m\}$ such that $\sigma_{a_m} \to \sigma_\infty$ and $\rho^*_{a_m} \to \rho^*_\infty$ as $m \to \infty$ for some $\sigma_\infty \in [0,\infty]$ and some $\rho^*_\infty \in \{\rho^* \in [0,1]^k : \|\rho^*\| \leq 1\}$. Then, by the continuity of $J(\sigma, \rho^*, V, c)$, we have with probability one,

$$J(\sigma_{a_m}, \rho^*_{a_m}, V_{a_m}, c) \to J(\sigma_\infty, \rho^*_\infty, V_\infty, c) \text{ as } m \to \infty. \tag{B.24}$$

Because the c.d.f. of $J(\sigma_\infty, \rho^*_\infty, V_\infty, c)$ is continuous and strictly increasing, we have

$$F^{-1}_{|J(\sigma_{a_m}, \rho^*_{a_m}, V_{a_m}, c)|}(1-\alpha) \to F^{-1}_{|J(\sigma_\infty, \rho^*_\infty, V_\infty, c)|}(1-\alpha). \tag{B.25}$$

Thus,

$$cv(1-\alpha, V_{a_m}, c) \to F^{-1}_{|J(\sigma_\infty, \rho^*_\infty, V_\infty, c)|}(1-\alpha) \leq cv(1-\alpha, V_\infty, c). \tag{B.26}$$

On the other hand, let $(\sigma_m^\dagger, \rho_m^{*\dagger})$ be such that $F^{-1}_{|J(\sigma_m^\dagger, \rho_m^{*\dagger}, V_\infty, c)|}(1-\alpha) \to cv(1-\alpha, V_\infty, c)$ as $m \to \infty$. Then,

$$\begin{aligned}
\liminf_{m\to\infty} cv(1-\alpha, V_{a_m}, c) &\geq \liminf_{m\to\infty} F^{-1}_{|J(\sigma^\dagger_{a_m}, \rho^{*,\dagger}_{a_m}, V_{a_m}, c)|}(1-\alpha) \\
&= F^{-1}_{|J(\sigma^\dagger_\infty, \rho^{*,\dagger}_\infty, V_\infty, c)|}(1-\alpha) \\
&= cv(1-\alpha, V_\infty, c),
\end{aligned} \tag{B.27}$$

where $(\sigma^\dagger_\infty, \rho^{*,\dagger}_\infty)$ is a cluster point of the sequence $\{(\sigma^\dagger_{a_m}, \rho^{*,\dagger}_m)\}_{m=1}^\infty$. Equations (B.26) and (B.27) together show (B.23).

As argued above, $\hat{V}_{a_n} \to_p V$. By the continuity of $cv(1-\alpha, \cdot, c)$ shown above, we have

$$cv(1-\alpha, \hat{V}_{a_n}, c) \to_p cv(1-\alpha, V, c). \tag{B.28}$$

Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr{}_{P_{a_n}} \left(|\hat{T}^{\text{mod}}_{a_n}(c)| > cv(1-\alpha, \hat{V}_{a_n}, c) \right) \\
&= \Pr\left(|J((\sigma, \rho^*, V, c)| > cv(1-\alpha, V, c) \right) \\
&\leq \Pr\left(|J(\sigma, \rho^*, V, c)| > F^{-1}_{|J(\sigma, \rho^*, V, c)|}(1-\alpha) \right) = \alpha.
\end{aligned} \tag{B.29}$$

□

Proof of Theorem 5.1. (a) Let $\widehat{LR1}_n$ and $\widehat{LR2}_n$ be the same as in (B.1). Then

$$\begin{aligned} n^{1/2}\widehat{LR1}_n &= n^{1/2}\sum\nolimits_{i=1}^{n}\left(\Lambda_i\left(\phi^*_{P_n}\right) - E_{P_n}\Lambda_i\left(\phi^*_{P_n}\right)\right) + n^{1/2}E_{P_n}\Lambda_i\left(\phi^*_{P_n}\right) \\ &\to_d \ \omega_{P_0}Z_\Lambda + d, \end{aligned} \tag{B.30}$$

by Lemma A.2(c) and Assumption 5.1(a). Also,

$$\begin{aligned} n^{1/2}\widehat{LR2}_n &= \frac{1}{\sqrt{n}}\left(n^{1/2}(\hat{\phi}_n - \phi^*_{P_n})\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\Lambda_i(\tilde{\phi}_n)}{\partial\phi\partial\phi'}\right)n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}) \\ &= n^{-1/2}\cdot O_p(1) \to_p 0, \end{aligned} \tag{B.31}$$

by Lemmas A.1 and A.2(a)-(c). Combining (B.30) and (B.31), we have

$$n^{1/2}\widehat{LR}_n \to_d \omega_{P_0}Z_\Lambda + d. \tag{B.32}$$

Let $W_{n,1}, W_{n,2}$ and $W_{n,3}$ be the same as in (B.3). Then

$$\begin{aligned} \hat{\omega}_n^2 = &-\widehat{LR}_n^2 + n^{-1}W_{n,1} + 2n^{-1}W_{n,2}\cdot n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}) \\ &+ n^{1/2}(\hat{\phi}_n - \phi^*_{P_n})'(n^{-1}W_{n,3})n^{1/2}(\hat{\phi}_n - \phi^*_{P_n}). \end{aligned} \tag{B.33}$$

Observe that, by the LLN, $n^{-1}W_{n,1} = n^{-1}\sum_{i=1}^{n}(\Lambda^*_{i,n})^2 \to_p \omega^2_{P_0}$. Also, similar to (B.5), we have $n^{-1}W_{n,2} = o_p(1)$, and similar to (B.6), we have $n^{-1}W_{n,3} = o_p(1)$. Thus,

$$\hat{\omega}_n^2 \to_p \omega^2_{P_0}. \tag{B.34}$$

By Lemma A.1 and Lemma A.2(a), we have $\hat{V}_n \to_p V_{P_0}$. Therefore,

$$\hat{T}_n^{\text{mod}} = \frac{n^{1/2}\widehat{LR}_n + n^{-1/2}tr(\hat{V}_n)/2}{[\hat{\omega}_n^2 + n^{-1}c\cdot tr(\hat{V}_n^2)]^{1/2}} \to_d \frac{\omega_{P_0}Z_\Lambda + d}{\omega_{P_0}} = Z_\Lambda + d/\omega_{P_0}, \tag{B.35}$$

which immediately implies the result of Theorem 5.1(a).

(b) The proof of part (b) follows the same steps as part (a) except $\widehat{LR}_n$ is now scaled by n instead of by $\sqrt{n}$ and $\hat{\omega}_n^2$ by n instead of by 1. Let λ_n stand for $nE_{P_n}\Lambda^*_{n,i}$. First,

$$\lambda_n^{-1}n\widehat{LR1}_n = \lambda_n^{-1}\sum\nolimits_{i=1}^{n}\left(\Lambda^*_{n,i} - E_{P_n}\Lambda^*_{n,i}\right) + 1 = o_p(1) + 1 \to_p 1, \tag{B.36}$$

where the second equality holds because

$$\lambda_n^{-2}E_{P_n}\left[\sum\nolimits_{i=1}^{n}\left(\Lambda^*_{n,i} - E_{P_n}\Lambda^*_{n,i}\right)\right]^2 = \lambda_n^{-2}n\omega^2_{P_n} \to 0\cdot\sigma^2_\infty = 0. \tag{B.37}$$

Similar to (B.31), we have $\lambda_n^{-1} n\widehat{LR2}_n = o_p(1)$. Thus,

$$\lambda_n^{-1} n\widehat{LR}_n \to_p 1. \tag{B.38}$$

Also, because $\lambda_n^{-2} E[\sum_{i=1}^n (\Lambda_{i,n}^*)^2] = \lambda_n^{-2} n\omega_{P_n}^2 + n^{-1} \to 0$,

$$\lambda_n^{-2} W_{n,1} \equiv \lambda_n^{-2} \sum\nolimits_{i=1}^n (\Lambda_{i,n}^*)^2 \to_p 0. \tag{B.39}$$

Similar to (B.5)-(B.6), we can show that $\lambda_n^{-2} W_{n,2} = o_p(1)$ and $\lambda_n^{-2} W_{n,3} = o_p(1)$. Therefore, by (B.3),

$$\lambda_n^{-2} n\hat{\omega}_n^2 \to_p 0. \tag{B.40}$$

By Lemma A.1 and Lemma A.2(a), we have $\hat{V}_n \to_p V_{P_0}$. We have shown in the proof of Theorem 4.1 that $cv(1-\alpha, V, c)$ is continuous in V. Thus, $cv(1-\alpha, \hat{V}_n, c) \to_p cv(1-\alpha, V_{P_0}, c)$. Then, by the continuous mapping theorem,

$$\lambda_n^{-1}[n\hat{\omega}_n^2 + c \cdot tr(\hat{V}^2)]^{1/2} \times cv(1-\alpha, \hat{V}_n, c) \to_p 0. \tag{B.41}$$

This implies that,

$$\begin{aligned}
&\Pr{}_{P_n}\left(\hat{T}_n^{\text{mod}}(c) > cv(1-\alpha, \hat{V}_n, c)\right) \\
&= \Pr\left(\lambda_n^{-1}\widehat{LR}_n + \lambda_n^{-1} tr(\hat{V}_n)/2 > \lambda_n^{-1}[n\hat{\omega}_n^2 + c \cdot tr(\hat{V}_n^2)]^{1/2} \cdot cv(1-\alpha, \hat{V}_n, c)\right) \\
&\to \Pr(1 > 0) = 1.
\end{aligned} \tag{B.42}$$

Similarly, one can show that $\Pr_{P_n}\left(\hat{T}_n^{\text{mod}}(c) < -cv(1-\alpha, \hat{V}_n, c)\right) \to \Pr(1 < 0) = 0$. □

References

Andrews, D. W. K. (1988): "Laws of Large Numbers for Dependent Non-identically Distributed Random Variables," *Econometric Theory*, 4, 458–467.

——— (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2009): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," unpublished manuscript, Cowles Foundation, Yale University.

Chen, X., H. Hong, and M. Shum (2007): "Nonparametric Likelihood Ratio Model Selection Tests between Parametric Likelihood and Moment Condition Models," *Journal of Econometrics*, 141, 109–140.

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

Cox, D. R. (1961): "Tests of Separate Families of Hypotheses," Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, University of California Press: Berkeley.

——— (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society. Series B (Methodological)*, 24, 406–424.

Gourieroux, C., and A. Monfort (1995): "Testing, Encompassing, and Simulating Dynamic Econometric Models," *Econometric Theory*, 11, 195–228.

Kitamura, Y. (2000): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," unpublished manuscript, Department of Economics, University of Pennsylvania.

——— (2006): "Empirical Likelihood Methods in Economics: Theory and Practice," Cowles Foundation Discussion Paper No. 1569, Cowles Foundation, Yale University.

Li, T. (2009): "Simulation Based Selection of Competing Structural Econometric Models," *Journal of Econometrics*, 148, 114–123.

Otsu, T., M. H. Seo, and Y.-J. Whang (2012): "Testing for Non-nested Conditional Moment Restrictions Using Unconditional Empirical Likelihood," *Journal of Econometrics*, 167, 370–382.

Otsu, T., and Y.-J. Whang (2011): "Testing for Non-nested Conditional Moment Retrictions via Conditional Empirical Likelihood," *Econometric Theory*, 27, 114–153.

Rivers, D., and Q. Vuong (2002): "Model Selection Tests for Nonlinear Dynamic Models," *Econometrics Journal*, 5, 1–39.

Schennach, S. M., and D. Wilhelm (2011): "A Simple Parametric Model Selection Test," unpublished manuscript, University of Chicago.

Shi, X. (2009): "Model Selection Tests for Nonnested Moment Inequality Models," unpublished manuscript, Department of Economics, Yale University.

Smith, R. J. (1997): "Alternative Semi-parametric Likelihood Approaches to Generalised Method of Moments Estimation," *Economic Journal*, 107, 503–19.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–33.

Online Supplemental Appendix for "A Nondegenerate Vuong Test"

Xiaoxia Shi
University of Wisconsin at Madison
xshi@ssc.wisc.edu

C Additional Simulation Example

In this section, I present the third simulation example. This example – borrowed from Schennach and Wilhelm (2011) – proves to be a very clever design that allows us to neatly demonstrate the power against both the $n^{-1/2}$-local alternatives defined in Assumption 5.1 and the n^{-1}-local alternatives defined in Assumption 5.2.

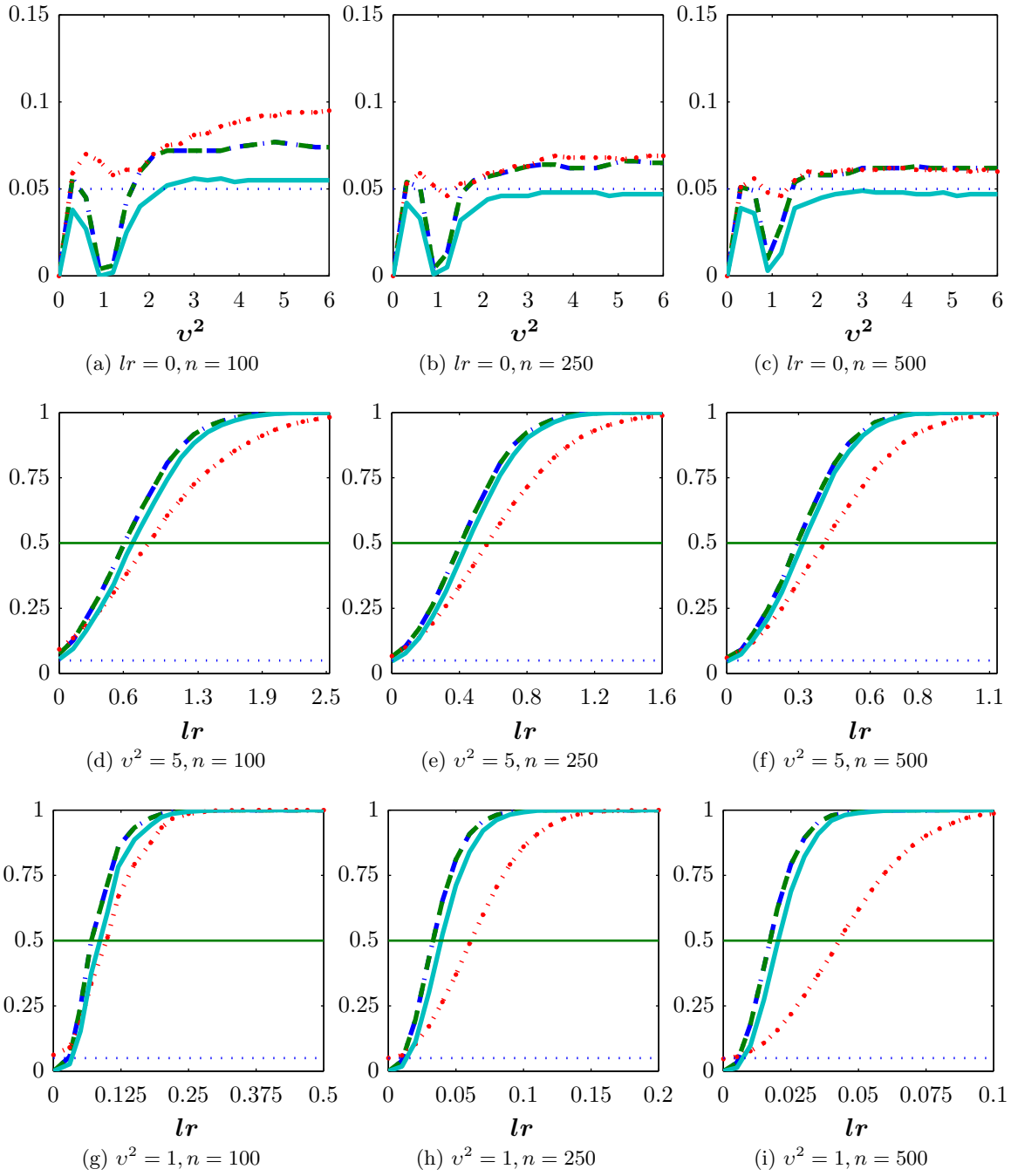
Example 3 (Normal Mean and Variance)**.** *Let the two models compared be:*
$\mathcal{F}$*:* $\{N(\theta,1):\theta\in\Theta\subset R\}$*, and*
$\mathcal{G}$*:* $\{N(0,\beta):\beta\in\mathcal{B}\subset(0,\infty)\}$.

Let Y be generated from $\sim N(\mu,v^2)$, where $\mu=\sqrt{e^{2\cdot lr-1+v^2}-v^2}$, where $lr\in\{x\in R: e^{2\cdot lr-1+v^2}-v^2\geq 0\}$. Under DGPs of this form, $E\left[\Lambda_i\left(\phi^*\right)\right]=lr$. Thus, varying lr controls how far the deviation is from H_0. On the other hand, when $lr=0$, varying the parameter v^2 controls how large ω^2 is. Setting $v^2=1$ makes $\omega^2=0$, and setting v^2 far from 1 makes ω^2 large.

First, I fix $lr=0$ and study the null rejection probabilities of the different tests. The simulation results are reported in the top three subplots of Figure 4 on the next page. The figure shows that my nondegenerate test has remarkable size control at all three sample sizes. On the other hand, the one-step and the two-step Vuong tests, as well as the SW tests have large size distortion at $n=100$, and still some noticeable size distortion at $n=250$.

Second, I fix $v^2=5$ and study the power of the different tests as lr varies from 0 to $1.6\sqrt{250/n}$. The $n^{-1/2}$-local power is considered because $v^2=5$ represents the nondegenerate case $\omega^2_{P_0}>0$, and local alternatives around this null DGP should have $\omega^2_{P_n}\nrightarrow 0$. The results are reported in the middle three subplots of Figure 4. The plots show that the power figures of all four tests stay constant as the sample size increases with $\sqrt{n}lr$ kept constant. My nondegenerate test has power similar to that of the one-step and the two-step Vuong tests and higher than that of the SW test. The power disadvantage of the SW test perhaps is due to the loss of efficiency from the sample splitting.

Last, I fix $v^2=1$ and study the power of the four different tests as lr varies from 0 to $0.2\cdot(250/n)$. The n^{-1}-local power is considered because $v^2=1$ represents the degenerate case: $\omega^2_{P_0}=0$. The results are reported in the middle three subplots of Figure 4. The plots show that the power of my nondegenerate test and that of the classical Vuong tests are similar and



(a) $lr = 0, n = 100$ (b) $lr = 0, n = 250$ (c) $lr = 0, n = 500$

(d) $v^2 = 5, n = 100$ (e) $v^2 = 5, n = 250$ (f) $v^2 = 5, n = 500$

(g) $v^2 = 1, n = 100$ (h) $v^2 = 1, n = 250$ (i) $v^2 = 1, n = 500$

Figure 4: Rejection probabilities for the one-step Vuong test (dash-dotted line), two-step Vuong test (dashed line), the SW test (dotted line) and my new nondegenerate test (solid line) for Example 3. The horizontal dotted line indicates the nominal level 5%.

all stay approximately constant as n increases with $n \times lr$ kept constant. On the other hand, the SW test appears to have lower power and its power decreases as n increases with $n \times lr$ kept constant. Thus, the power gap between my new nondegenerate test and the SW test increases as the sample size increases.

D Schennach and Wilhelm (2011) Test

Here I briefly describe Schennach and Wilhelm's (2011) split sample test (SW test) in my notation. To construct the SW test, first, split the full sample $\{X_i\}_{i=1}^n$ into two equal-sized samples $\{X_{(1)i}\}_{i=1}^{n/2}$ and $\{X_{(2)i}\}_{i=1}^{n/2}$ (for example, split into two halves according to the natural ordering); second, let the split-sample log-likelihood ratio estimator be

$$\begin{aligned}\widehat{LR}_n^{splt} =& \frac{2}{n(2+\varepsilon_n)} \sum\nolimits_{i=1}^{n/2} \Big[\Big(\log f(X_{(1)i}, \hat{\theta}_n) - \log g(X_{(2)i}, \hat{\beta}_n) \Big) \\ & + (1+\varepsilon_n) \Big(\log f(X_{(2)i}, \hat{\theta}_n) - \log g(X_{(1)i}, \hat{\beta}_n) \Big) \Big], \end{aligned} \tag{D.1}$$

where $\varepsilon_n \in R \backslash \{0, -2\}$ is a user-chosen weighting. Let the variance estimator be

$$\begin{aligned}(\hat{\omega}_n^{splt})^2 =& \frac{2}{n(2+\varepsilon_n)^2} \sum\nolimits_{i=1}^{n/2} \Big[\Big(\log f(X_{(1)i}, \hat{\theta}_n) - \log g(X_{(2)i}, \hat{\beta}_n) \Big) \\ & + (1+\varepsilon_n) \Big(\log f(X_{(2)i}, \hat{\theta}_n) - \log g(X_{(1)i}, \hat{\beta}_n) \Big) \Big]^2 - (\widehat{LR}_n^{splt})^2; \text{ and} \end{aligned} \tag{D.2}$$

third, let

$$\hat{T}_n^{splt} = (n/2)^{1/2} \widehat{LR}_n^{splt} / \hat{\omega}_n^{splt}. \tag{D.3}$$

Finally, reject H_0 if $|\hat{T}_n^{splt}| > z_{\alpha/2}$. When H_0 is rejected, pick model $\mathcal{F}$ if $\widehat{LR}_n^{splt} > 0$ and pick model $\mathcal{G}$ if $\widehat{LR}_n^{splt} < 0$.[15]

In the 2011 version of their paper, they suggest a robust choice for the weighting parameter ε_n:

$$\varepsilon_n^* = \max \left\{ \frac{Cov_n(\log f_i(\hat{\theta}_n), \log g_i(\hat{\beta}_n))}{Var_n(\log f_i(\hat{\theta}_n)) + Var_n(\log g_i(\hat{\beta}_n))}, 0 \right\} - 1, \tag{D.4}$$

where Cov_n stands for sample covariance and Var_n stands for sample variance.

[15] Schennach and Wilhelm (2011) write their test as a Wald test from a GMM problem formed by the MLE f.o.c.s and the null hypothesis of the Vuong test. Some algebra shows that their GMM estimators of θ and β are exactly the maximum likelihood estimators because they have to satisfy the MLE f.o.c.s, and their regularized Wald statistic is exactly $(\hat{T}_n^{splt})^2$.