

# A Simple Distribution-Free Test for Nonnested Hypotheses

Kevin A. Clarke<sup>†</sup>  
University of Rochester

## Abstract

In this paper, we more fully develop the properties of the distribution-free test for nonnested model discrimination introduced by Clarke (2003). We prove that the test is both consistent and unbiased. We demonstrate that the test is asymptotically more efficient for highly leptokurtic distributions than the well-known Vuong test. Using a Monte Carlo experiment, we then establish that the distribution of individual log-likelihood ratios (the data to which both tests are applied) is highly leptokurtic. Finally, we use the same Monte Carlo to measure the performance of the distribution-free test and the Vuong test. The Monte Carlo advances previous efforts in that it allows for two misspecified models that vary in distance from a true, but “unknown,” data generating process. The results indicate that the power of the new test is as great as or, for many alternatives, significantly greater than the power of the Vuong test.

June 15, 2004

---

<sup>†</sup>Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: [kevin.clarke@rochester.edu](mailto:kevin.clarke@rochester.edu). Previous versions of this paper were presented at the Midwest Political Science Association Meetings, Chicago IL, April 2004 and the NorthEast Methodology Program, New York University, 2003; I thank the participants for their comments. Support from the National Science Foundation (Grant #SES-0213771) is gratefully acknowledged.

# 1 Introduction

The purpose of this paper is to more fully develop the properties of the distribution-free test for nonnested model discrimination proposed in Clarke (2003). To that end, we build upon the results of Clarke (2003) in four significant ways. First, we prove two basic properties that any hypothesis test should possess, unbiasedness and consistency. Second, while the previous paper established the greater power of the distribution-free test over the well-known Vuong test, the current paper goes further and demonstrates *why* the distribution-free test has greater power. The answer lies in the shape of the data to which the tests are applied. Asymptotic relative efficiency calculations show that the distribution-free test has greater power than the Vuong test for highly leptokurtic distributions (those with taller peaks and heavy tails than the normal). Results from a Monte Carlo experiment demonstrate that the distribution of individual log-likelihood ratios, the data to which model selection tests are applied, is indeed highly leptokurtic. We therefore expect the distribution-free test to have greater power than the Vuong test.

The third improvement concerns the Monte Carlo itself. The experiment in Clarke (2003) compares two models, one of which is the data generating process. The Monte Carlo experiment in the current paper builds on those results by comparing two models that vary in distance from a true, but “unknown,” data generating process. Given that a researcher is rarely lucky enough to compare two models, one of which is true, this approach more accurately reflects the situation in which substantive researchers find themselves. The results confirm our expectation in that the distribution-free test has significantly greater power than the Vuong test.

Finally, although we argue that model selection tests are superior to model selection criteria, the latter see considerable use in political science. Some results comparing the distribution-free test to model selection criteria are therefore presented.

As we are concerned in this paper with the technical properties of the distribution-free test and the Vuong test, we direct those readers interested in substantive applications of these tests to Clarke (2003) and Clarke and Signorino (2004).

## 2 Why Model Selection Tests

There are four common approaches to the problem of discriminating among nonnested hypotheses: tests of absolute discrimination (such as the Cox test), model selection tests (such as the Vuong test), Bayesian methods (such as the posterior odds ratio and Bayes Factors), and model selection criteria (such as Akaike's information criteria (AIC) and Schwarz's Bayesian information criteria (BIC)). Why we focus on model selection tests, as opposed to the other approaches mentioned above, is a reasonable question. In deciding upon a particular approach to nonnested model discrimination, two sets of criteria must be kept in mind.

One set comprises two technical properties relevant mostly to methodologists. These criteria are unbiasedness and consistency. A test is unbiased if the probability of rejecting a false null is greater than the probability of rejecting a true null. A test is consistent if the probability of rejecting a false null goes to one asymptotically. These properties are often the minimal requirements placed on the power function of any test, and consequently, most practical tests meet these conditions.

The other set comprises the practical concerns of substantive researchers. These additional criteria include ease of calculation and ease of interpretation. Methods that do not require simulations or bootstrapping are preferred to those that do. By the same token, methods that can be easily performed in mainstream statistical software are preferred to those methods that cannot. Finally, methods that indicate the strength of the evidence in favor of a particular model and convey a clear understanding of the uncertainty regarding model choice are preferred to those that do not. Unsurprisingly, it is on this second set of criteria that most approaches to nonnested model discrimination fail.<sup>1</sup>

The Cox test (Cox 1961), for instance, is difficult to calculate for specifications more complex than ordinary least squares regression. The most successful approach to calculating the test for nonlinear specifications requires simulations to estimate the pseudo-true values used in the calculation (Pesaran and Pesaran 1993). The Cox test is also difficult to interpret in many circumstances. The reason lies in the fact that the Cox test is a test of absolute discrimination where each model is evaluated against the data with

---

<sup>1</sup>The above should not be interpreted to mean that there are no circumstances under which these other approaches might prove useful. AIC and BIC, for instance, have proven useful in time-series analysis.

the alternative model providing power.<sup>2</sup> The problem with absolute tests is that both models may be “accepted” or rejected leading to an ambiguous outcome (Granger, King, and White 1995). In such a case, the test does not provide evidence in favor of a particular model thus violating another of our criteria. Finally, Monte Carlo simulations demonstrate that the test lacks power in many commonly encountered situations (Clarke 2001).<sup>3</sup>

The Bayesian approaches fare only slightly better. The posterior odds ratio is difficult to calculate and has yet to see any significant use in political science. Bayes factors have seen more use in political science than posterior odds ratios (see Smith 1999 and Bartels 1997), but the ease of their calculation varies with the approximation used. The Laplace approximation (Kass and Raftery 1995) and the data augmentation approach (Albert and Chib 1993) are accurate, but difficult and time-consuming to calculate. The BIC approximation (Raftery 1995) is quite easy to calculate, but less accurate. In either case, interpreting a Bayes factor is not a straightforward exercise. First, there is no natural metric for interpreting the size of a Bayes factor, although “rules of thumb” have been suggested (Raftery 1995). Second, the Bayes factor does not provide a measure of support for one model over another, but rather it measures “the change in the odds in favor of the hypothesis when going from prior to the posterior” (Lavine and Schervish 1999). That is, a small Bayes factor may indicate only that the data lower the prior probability of a hypothesis, not that the hypothesis is unlikely. Thus, Bayes factors do not indicate the strength of the evidence in favor of a particular model. This point is widely misunderstood.

Model selection criteria and model selection tests are easy to calculate and interpret, and thus do not suffer from the same drawbacks as the Cox test and the Bayesian approaches. The main difference between the criteria and tests lies in the probabilistic framework of the latter (Amemiya 1980; Vuong 1989). Whereas a model selection criterion always chooses a “best” model (the model that minimizes the criterion), the distributional results associated with the hypothesis tests allow the possibility of concluding that there is insufficient evidence to reject a conclusion of equivalence. The ability to choose neither model is a significant advantage of the hypothesis tests, but it is not determinant, and we present some comparative results at the end of

---

<sup>2</sup>Model selection tests are tests of relative discrimination where the models are evaluated against the data and each other.

<sup>3</sup>The concepts of power and size must be modified when applied to nonnested hypothesis testing; see Pesaran (1974).

the paper. For the moment, we focus on determining the conditions under which we prefer either the Vuong test or the distribution-free test.

### 3 The Vuong and Distribution-Free Tests

The Vuong test (Vuong 1989) and the distribution-free test introduced by Clarke (2003) are based on the Kullback-Leibler information criteria (Kullback and Leibler 1951), a measure of the “distance” between two statistical models.<sup>4</sup> Vuong (1989) defines the KLIC as,

$$\text{KLIC} \equiv E_0[\ln h_0(Y_i|X_i)] - E_0[\ln f(Y_i|X_i; \beta_*)], \quad (1)$$

where  $h_0(\cdot|\cdot)$  is the true conditional density of  $Y_i$  given  $X_i$  (that is, the true but unknown model),  $E_0$  is the expectation under the true model, and  $\beta_*$  are the pseudo-true values of  $\beta$  (the estimates of  $\beta$  when  $f(Y_i|X_i)$  is not the true model).

The best model is the model that minimizes the equation given above. That is, the best model is the one that is closest to the true, but unknown, specification. The model that is closest to the true specification must therefore be the model that maximizes  $E_0[\ln f(Y_i|X_i; \beta_*)]$ .<sup>5</sup> In other words, when comparing two rival models, we should select one model over the other if the individual log-likelihoods of the first model are significantly larger than the individual log-likelihoods of its rival.

#### 3.1 The Vuong Test

The null hypothesis of Vuong’s test is,

$$H_0 : E_0 \left[ \ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} \right] = 0, \quad (2)$$

which states that two rival models are equally close to the true specification.<sup>6</sup>

---

<sup>4</sup>“Distance” is in quotes as the KLIC between models  $f$  and  $g$  is generally not equal to the KLIC between models  $g$  and  $f$ .

<sup>5</sup>To understand how it is possible to know which model maximizes the expected value under an unknown specification, see White (1994) on quasi-maximum likelihood estimators.

<sup>6</sup> $\gamma_*$  and  $Z_i$  in model  $g$  are analogous to  $\beta_*$  and  $X_i$  in model  $f$ .

Because the true model is unknown, the expected value in the above hypothesis is unknown. Vuong proves that under general conditions,

$$\frac{1}{n}LR_n(\hat{\beta}_n, \hat{\gamma}_n) \xrightarrow{a.s.} E_0 \left[ \ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} \right], \quad (3)$$

which states that the expected value given in the null hypothesis can be consistently estimated by  $(\frac{1}{n})$  times the likelihood ratio statistic. The actual test is then,

$$\text{under } H_0 : \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0, 1), \quad (4)$$

where

$$LR_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv L_n^f(\hat{\beta}_n) - L_n^g(\hat{\gamma}_n) \quad (5)$$

and

$$\hat{\omega}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n \left[ \ln \frac{f(Y_i|X_i; \hat{\beta}_n)}{g(Y_i|Z_i; \hat{\gamma}_n)} \right]^2 - \left[ \frac{1}{n} \sum_{i=1}^n \ln \frac{f(Y_i|X_i; \hat{\beta}_n)}{g(Y_i|Z_i; \hat{\gamma}_n)} \right]^2. \quad (6)$$

The Vuong test can be described in simple terms. If the null hypothesis is true, the average value of the individual log-likelihood ratios should be zero.<sup>7</sup> If  $H_f$  is true, the average value of the individual log-likelihood ratios should be significantly greater than zero. If the reverse is true, the average value of the individual log-likelihood ratios should be significantly less than zero. In other words, the Vuong test statistic is simply the average log-likelihood ratio suitably normalized.<sup>8</sup>

The Vuong statistic is sensitive to the number of estimated coefficients in each model, and therefore the test must be corrected for the dimensionality of the models. Vuong (1989) suggests using a correction that corresponds to either Akaike's (1973) information criteria or Schwarz's (1978) Bayesian information criteria. If using the latter, the adjusted statistic becomes,

---

<sup>7</sup>Recall that the log-likelihood reported by statistical software is the sum of the log-likelihoods for each individual observation. As a log-likelihood ratio is simply the difference between two log-likelihoods, there are  $n$  individual log-likelihood ratios — one for each observation.

<sup>8</sup>R code for the Vuong test, the distribution-free test, BIC, and AIC, for any of R's GLM commands, is available from the authors.

$$L\tilde{R}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv LR_n(\hat{\beta}_n, \hat{\gamma}_n) - \left[ \left(\frac{p}{2}\right) \ln n - \left(\frac{q}{2}\right) \ln n \right], \quad (7)$$

where  $p$  and  $q$  are the number of estimated coefficients in models  $f$  and  $g$ , respectively.

### 3.2 The Distribution-Free Test

Clarke’s (2003) distribution-free test applies a modified paired sign test to the differences in the individual log-likelihoods from two nonnested models. Whereas the Vuong test determines whether or not the average log-likelihood ratio is statistically different from zero, the proposed test determines whether or not the *median* log-likelihood ratio is statistically different from zero. If the models are equally close to the true specification, half the individual log-likelihood ratios should be greater than zero and half should be less than zero. If model  $f$  is “better” than model  $g$ , more than half the individual log-likelihood ratios should be greater than zero. Conversely, if model  $g$  is “better” than model  $f$ , more than half the individual log-likelihood ratios should be less than zero.

Utilizing Vuong’s notation, the null hypothesis of the distribution-free test is,

$$H_0 : \text{median}_0 \left[ \ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} \right] = 0 \quad (8)$$

or, equivalently,

$$H_0 : \Pr_0 \left[ \ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)} > 0 \right] = .5. \quad (9)$$

The two assumptions of the test are unsurprising and quite general. First, the differences,  $\ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)}$ , are mutually independent.<sup>9</sup> Second, each  $\ln \frac{f(Y_i|X_i; \beta_*)}{g(Y_i|Z_i; \gamma_*)}$  comes from a continuous population (not necessarily the same) that has a common median  $\theta$ . Proofs of consistency and unbiasedness for the distribution-free test are in the Appendix.

---

<sup>9</sup>This assumption does not mean that the individual log-likelihoods themselves must be independent, only that the ratios (the differences in the individual log-likelihoods) be mutually independent.

Letting  $d_i = \ln f(Y_i|X_i; \hat{\beta}_n) - \ln g(Y_i|Z_i; \hat{\gamma}_n)$ , the test statistic is,

$$B = \sum_{i=1}^n I_{(0,+\infty)}(d_i). \quad (10)$$

The test statistic is simply the number of positive differences, and it is distributed Binomial with parameters  $n$  and  $\theta = 0.5$ .<sup>10</sup>

If model  $f$  is “better” than model  $g$ ,  $B$  will be significantly larger than its expected value under the null hypothesis ( $n/2$ ). For an upper tail test, we reject the null hypothesis of equivalence when  $B \geq c_\alpha$ , where  $c_\alpha$  is chosen to be the smallest integer such that,

$$\sum_{c=c_\alpha}^n \binom{n}{c} 0.5^n \leq \alpha. \quad (11)$$

For a lower tail test, the inequality is reversed, and the sum goes from  $c = 0$  to  $c = c_\alpha$ .

One of the great strengths of this procedure is that its implementation is remarkably simple; the test can be produced by any mainstream statistical software package using the following algorithm:

1. Run model  $f$ , saving the individual log-likelihoods,  $\ln f(Y_i|X_i; \hat{\beta}_n)$ .
2. Run model  $g$ , saving the individual log-likelihoods,  $\ln g(Y_i|Z_i; \hat{\gamma}_n)$ .
3. Compute the differences,  $d_i$ , and count the number of positive values.
4. The number of positive differences,  $B$ , is distributed Binomial( $n, .5$ ).

This test, like the Vuong test, is sensitive to the dimensionality of the competing models. Once again, we need a correction for the degrees of freedom. The Schwarz correction is,

$$\left[ \left( \frac{p}{2} \right) \ln n - \left( \frac{q}{2} \right) \ln n \right], \quad (12)$$

---

<sup>10</sup>Just as in the paired sign test, we test whether the rival models are different by some non-zero constant,  $C$ , by simply subtracting  $C$  from each of the differences and then computing the test statistic (Bradley 1968).

<sup>11</sup>Because the test statistic is discreet,  $\alpha$  often cannot be achieved exactly. See Section 4.3 for a complete discussion.

where  $p$  and  $q$  are the number of estimated coefficients in models  $f$  and  $g$ , respectively.

As we are working with the individual log-likelihood ratios, we cannot apply this correction to the “summed” log-likelihood ratio as Vuong did for his test. We can, however, apply the *average* correction to the individual log-likelihood ratios. That is, we correct the individual log-likelihoods for model  $f$  by a factor of,

$$\left(\frac{p}{2n}\right) \ln n, \tag{13}$$

and the individual log-likelihoods for model  $g$  by a factor of,

$$\left(\frac{q}{2n}\right) \ln n. \tag{14}$$

While we cannot justify any particular correction, we can broadly justify the approach by appealing to Vuong’s justification for his correction. Vuong notes that as long as the correction factor divided by the square root of  $n$  has a stochastic order of 1,

$$n^{-1/2}K_n(\mathbf{F}_\theta, \mathbf{G}_\gamma) = o_p(1), \tag{15}$$

the adjusted statistic has the same asymptotic properties of the unadjusted statistic.

Vuong’s justification amounts to pointing out that the asymptotic properties of the adjusted statistic are the same as the asymptotic properties of the unadjusted statistic. If we consider the normal approximation to the distribution-free test (see Section A.1 of the Appendix), we can see that the asymptotic properties of the distribution-free test are also unaffected by the correction.

### 3.3 Comparing the Tests

Given that we have competing tests of the same general hypothesis that both meet the criteria laid out in Section 2, we need a method of comparing the tests beyond that offered to us by Monte Carlo experiments.<sup>12</sup> One such method is known as Pitman efficiency or asymptotic relative efficiency. The

---

<sup>12</sup>Some of the results given in this section might normally be found in an appendix. We feel, however, that a more leisurely exposition is appropriate given the unfamiliarity many political scientists will have with this material.

criterion is credited to Pitman, in an unpublished paper, with generalizations by Noether (1955) and Hodges and Lehmann (1956).

Noether (1967) provides the following definition of Pitman efficiency,

If we have two tests of the same hypothesis and the same significance level and if for the same power with respect to the same alternative one test requires a sample of size  $N_1$  and the other a sample of size  $N_2$ , then the relative efficiency of the first test with respect to the second test is given by the ratio  $e = e_{1,2} = N_2/N_1$ .

This definition, as Noether notes, is problematic in most cases as the efficiency of a test depends on the significance level  $\alpha$ , the alternative  $\theta$ , and the sample size of the first test  $N_1$ . In recognition of this dependence,  $e$  should be written  $e(\alpha, \theta, N)$ . The complication of evaluating the relative efficiency of two tests in terms of the three arguments  $\alpha$ ,  $\theta$ , and  $N$ , however, may be avoided by consideration of the limiting case, or asymptotic relative efficiency, which does not depend on  $\alpha$ . That is, we look at the ratio of sample sizes,  $N_2/N_1$ , as  $N \rightarrow \infty$ , or,

$$e = \lim_{N \rightarrow \infty} e(\alpha, \theta, N). \quad (16)$$

For consistent tests, such as the Vuong and distribution-free, the power approaches 1 for large sample sizes with fixed alternatives. To avoid comparing two tests whose power is 1, we let the alternative hypothesis  $\theta$  approach the null hypothesis  $\theta_0$  so that the power of each lies on the open interval  $(\alpha, 1)$  (Gibbons and Chakraborti 1992). The asymptotic relative efficiency (sometimes known as local asymptotic efficiency) is then,

$$e = \lim_{\substack{N \rightarrow \infty \\ \theta \rightarrow \theta_0}} e(\alpha, \theta, N). \quad (17)$$

To further lessen the computational burden, Pitman showed that under certain regularity conditions, the asymptotic relative efficiency of one test with respect to another test is equal to the limit of the ratio of efficacies. That is, given two tests  $T$  and  $T^*$ ,

$$\text{A.R.E.}(T, T^*) = \lim_{n \rightarrow \infty} \frac{\text{eff}(T_n)}{\text{eff}(T_n^*)}, \quad (18)$$

where  $\text{eff}(T_n)$  is the efficacy of the test statistic  $T_n$  for the hypothesis  $\theta = \theta_0$ ,

$$\text{eff}(T_n) = \frac{[dE(T_n)/d\theta]^2|_{\theta=\theta_0}}{\text{Var}(T_n)|_{\theta=\theta_0}}. \quad (19)$$

Proof of the equivalence between the limiting efficacy ratio and asymptotic relative efficiency, along with the regularity conditions, are given by Gibbons and Chakraborti (1992).<sup>13</sup>

As noted in Section 3.2, the distribution-free test is based on the paired sign test, the efficacy of which is a standard result given by Noether (1967), Randles and Wolfe (1979), and Gibbons and Chakraborti (1992). For  $N$  observations from any population  $F_D$  (where  $D = X - Y$ ) with median  $\theta$ , the efficacy is,

$$\begin{aligned} \text{eff}(B_n) &= 4N f_D^2(\theta) \\ &= 4N f^2[F^{-1}(0.5)]. \end{aligned} \quad (20)$$

For a normal distribution,  $F_X \sim N(\theta, \sigma^2)$ ,  $f_X$  reduces to,

$$f_X = \frac{1}{\sigma\sqrt{2\pi}}, \quad (21)$$

and the efficacy is,

$$\begin{aligned} \text{eff}(B_n) &= 4N f^2[F^{-1}(0.5)] \\ &= 4N * \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^2 \\ &= \frac{2N}{\pi\sigma^2}. \end{aligned} \quad (22)$$

For a double exponential (or Laplace) distribution<sup>14</sup>,  $f_X$  reduces to,

$$f_X = \frac{1}{2\lambda}, \quad (23)$$

and the efficacy is,

---

<sup>13</sup>The regularity conditions are quite general, and both tests under consideration here meet the conditions.

<sup>14</sup> $f_X(x) = \frac{1}{2\lambda}e^{-|x-\mu|/\lambda}$ , where  $E[X] = \mu$ , and  $\text{Var}[X] = 2\lambda^2$ . The reason for choosing the double exponential will become apparent.

$$\begin{aligned}
\text{eff}(B_n) &= 4Nf^2[F^{-1}(0.5)] \\
&= 4N * \left(\frac{1}{2\lambda}\right)^2 \\
&= \frac{N}{\lambda^2}.
\end{aligned} \tag{24}$$

The efficacy of the Vuong test is not difficult to calculate. We can write the test statistic in the following way:

$$\begin{aligned}
V_n = \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{N})\hat{\omega}_n} &= \frac{\frac{1}{N}LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{\hat{\omega}_n/\sqrt{N}} \\
&= \frac{\sqrt{N}[\frac{1}{n}LR_n(\hat{\beta}_n, \hat{\gamma}_n)]}{\hat{\omega}_n}.
\end{aligned} \tag{25}$$

Letting  $\bar{D} = \frac{1}{n}LR_n(\hat{\beta}_n, \hat{\gamma}_n)$  and  $\mu_D = E_0 \left[ \ln \frac{f(Y_t|X_t;\beta_*)}{g(Y_t|Z_t;\gamma_*)} \right]$ , we see that the above is equal to,

$$\left[ \frac{\sqrt{N}(\bar{D} - \mu_D)}{\omega} + \frac{\sqrt{N}\mu_D}{\omega} \right] \frac{\omega}{\hat{\omega}}. \tag{26}$$

Given that  $\lim_{N \rightarrow \infty} (\hat{\omega}_n/\omega) = 1$  (see Vuong (1989), p. 315), the expected value and variance of the Vuong test statistic for large  $N$  are,

$$\begin{aligned}
E[V_n] &= \frac{\sqrt{N}\mu_D}{\omega}, \\
\text{Var}[V_n] &= \frac{N \text{Var}(\bar{D})}{\omega^2} = 1.
\end{aligned} \tag{27}$$

The efficacy is therefore,

$$\begin{aligned}
\text{eff}(V_n) &= \frac{[dE(V_n)/d\theta]^2|_{\mu_D=0}}{\text{Var}(V_n)|_{\mu_D=0}} \\
&= \frac{[\sqrt{N}/\omega]^2}{1} \\
&= \frac{N}{\omega^2}.
\end{aligned} \tag{28}$$

For a normal distribution,  $F_X \sim N(\theta, \sigma^2)$ , the efficacy of the Vuong is simply  $N/\sigma^2$ . For the double exponential, the efficacy is,

$$\text{eff}(V) = \frac{N}{\omega^2} = \frac{N}{2\lambda^2}. \quad (29)$$

We are now in a position to make some statements regarding the asymptotic relative efficiency of the Vuong test versus the distribution-free test. For normally distributed data, the asymptotic relative efficiency is,

$$\begin{aligned} \text{A.R.E.}(B, V) &= \lim_{n \rightarrow \infty} \frac{\text{eff}(B_n)}{\text{eff}(V_n)} \\ &= \frac{2N/\pi\sigma^2}{N/\sigma^2} \\ &= \frac{2}{\pi}. \end{aligned} \quad (30)$$

This result means that if the distribution of individual log-likelihood ratios is normal, the distribution-free test is only  $2/\pi = 0.637$  or 64% as efficient as the Vuong test. The distribution-free test would be even more inefficient for platykurtic distributions ( $\kappa < 3$ ). Under such conditions, we are better off using the Vuong test as it provides greater power than the distribution-free test.

Things look quite different, however, when we consider highly leptokurtic (heavy-tailed and high peaked) distributions such as the double exponential. For data that is distributed according to the double exponential, the asymptotic relative efficiency is,

$$\begin{aligned} \text{A.R.E.}(B, V) &= \lim_{n \rightarrow \infty} \frac{\text{eff}(B_n)}{\text{eff}(V_n)} \\ &= \frac{N/\lambda^2}{N/2\lambda^2} \\ &= 2. \end{aligned} \quad (31)$$

This result means that if the individual log-likelihood ratios are distributed double exponential, the Vuong test is only 50% as efficient as the distribution-free test. The Vuong would be even more inefficient for distributions such as

the Cauchy, which has extremely heavy tails. Under these conditions, we are better off using the distribution-free test as it provides greater power than the Vuong test.

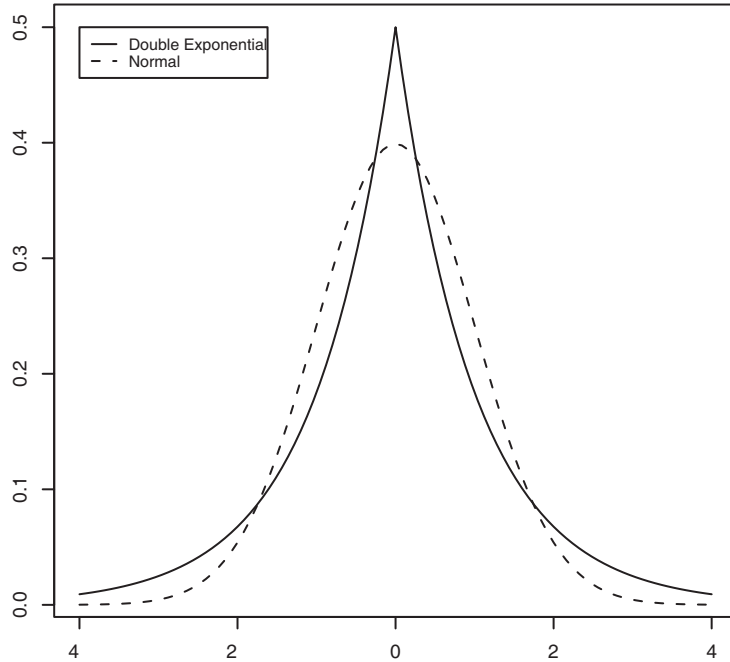


Figure 1: The Double Exponential and the Normal

The question now facing us is whether the distribution of individual log-likelihoods ratios across different models and functional forms looks more like a normal distribution or more like a double exponential distribution. Figure 1 shows the difference between these distributions. The double exponential distribution has heavier tails and is more “peaked” than the normal distribution. For such distributions, the distribution-free test has greater power than the Vuong test. For distributions with lighter tails and smaller “peaks”, the Vuong has greater power.

We can measure the tail weight and peakedness of a distribution using the kurtosis coefficient (Spanos 1999). Defined by Balanda and MacGillivray (1988) as “the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails,” the common measure of kurtosis is the standardized fourth central moment,

$$\kappa = \frac{\mu_4}{(\mu_2)^2}, \quad (32)$$

where,

$$\mu_r(\boldsymbol{\theta}) = E[X^r] = \int_{-\infty}^{\infty} (x - \mu)^r f(x; \boldsymbol{\theta}) dx, \quad r = 2, 3, \dots \quad (33)$$

The kurtosis coefficient for a normal distribution is 3 and for a double exponential 6. As the Monte Carlo experiment in the next section demonstrates, the distribution of individual log-likelihood ratios is much closer to being a double exponential distribution than a normal distribution. We therefore expect the distribution-free test to have greater power than the Vuong.

## 4 The Monte Carlo Experiment

The Monte Carlo experiment performed in Clarke (2003) compared two rival probit models, one of which served as the data generating process (DGP).<sup>15</sup> The experiment was overly-stylized in the sense that it is highly unlikely that a researcher is ever lucky enough to compare two models, one of which is true. It is far more likely that both models are misspecified in some fashion, and that the true model is unknown. The innovation in the experiment presented in this paper is the specification of three models, one of which is the DGP and two that vary in distance from each other and the DGP.<sup>16</sup>

### 4.1 Set-up of the Experiment

In each replication, six variables with zero means and unit variances are drawn from a multivariate normal distribution. The first two variables are used to form the DGP (along with a randomly-drawn normally distributed error term with a standard deviation that varies). Each of the other two sets of variables are used to form the two rival models. The six variables are drawn with a given correlation matrix that controls the canonical correlations between the three models as well as the bivariate correlations within the models (see Kaiser and Dickman 1962).

---

<sup>15</sup>The probit was chosen for convenience and for its ubiquity in international relations research.

<sup>16</sup>R code for replicating the experiment reported here is available from the authors.

Following Davidson and MacKinnon (1993), we are interested in tests in regression directions; that is, tests of the specification of the regression function. To that end, other than the size of the sample and the signal-to-noise ratio (the variance of the systematic portion of the model to that of the stochastic portion), the only variation in the experiment is the distance of the alternative hypothesis from the null hypothesis. Let the two models that do not serve as the DGP be models  $f$  and  $g$ . The canonical correlation between model  $g$  and the DGP is set at 0.2. The canonical correlation between model  $f$  and the DGP varies from 0.3 to 0.9. Therefore, the alternative hypothesis is closest to the null when the canonical correlation between model  $f$  and the DGP is at 0.3, and farthest from the null when the canonical correlation is at 0.9.<sup>17</sup>

The experiment was performed with the following parameters:

1.  $n = (50, 100, 200, 500, 1000)$
2. Distance of the alternative from the null = (.3, .4, .5, .6, .7, .8, .9)
3. Error standard deviation = (1, 2)
4. Tests = Vuong, Distribution-free, BIC.

Thus, 70 variations on the experiment were performed. Each replication in each variation led to either a rejection or acceptance of the null hypothesis for both tests, as well as a decision based on the minimum BIC (see Section 4.5). We can therefore treat each replication as an independent Bernoulli trial and use the obvious estimator of power, the number of rejections over the number of replications (Davidson and MacKinnon 1993).<sup>18</sup> 8000 replications of each variation were run to ensure that the 95% confidence interval on each estimate is approximately 0.01. All coefficients are set to 1, and the 6 independent variables, as well as the error term, were drawn anew for each replication.<sup>19</sup>

---

<sup>17</sup>In the discussion that follows, the distance between the null and the alternative is indexed only by this correlation.

<sup>18</sup>This estimator is not the efficient estimator. Less variable estimators can be designed using control variates (see Hendry (1984) for more information). Given that the experiment is not overly time consuming, a large number of replications was possible thus making efficiency a less pressing concern.

<sup>19</sup>A new set of variables is drawn for each replication to avoid results that may depend on the idiosyncratic characteristics of a particular draw. See Davidson and MacKinnon (1993) for a discussion of the pros and cons.

## 4.2 Simulation Results 1: Kurtosis

Along with a decision for both tests, the experiment reported the kurtosis coefficient for the empirical distribution of the individual log-likelihood ratios for each replication. Table 1 shows the average kurtosis value for all sample sizes and alternatives in the experiment.

Distance	Sample Size				
	50	100	200	500	1000
3	5.19	5.44	5.49	5.50	5.49
4	5.26	5.52	5.58	5.57	5.59
5	5.24	5.49	5.62	5.68	5.73
6	5.32	5.62	5.81	5.88	5.90
7	5.40	5.70	5.89	6.09	6.14
8	5.47	5.87	6.10	6.34	6.38
9	5.58	5.98	6.35	6.58	6.70

Table 1: Mean Kurtosis Coefficient of the Empirical Distribution of Individual Log-Likelihood Ratios

The results range from a kurtosis coefficient of 5.2 for a sample size of 50 and an alternative near the null to a kurtosis coefficient of 6.7 for a sample size of 1000 and an alternative far from the null. The distribution of individual log-likelihood ratios is therefore unlikely to be normally distributed given that the normal has a kurtosis of 3.

Of course, kurtosis does not fully characterize the shape of a distribution. It is possible for two distributions with the same kurtosis coefficient to have vastly different shapes. In this case, however, the distribution does indeed have a higher peak and heavier tails than the normal. That is, it looks more like a double exponential distribution. Figure 2 shows one such representative distribution with the normal and double exponential distributions superimposed. The closer match to the double exponential is clear.<sup>20</sup> While the functional form of the rival models can affect the exact shape of this distribution, heavier tails and higher peaks are characteristic.

Given that the distribution of individual log-likelihood ratios looks more like a double exponential than a normal, we expect the distribution-free

<sup>20</sup>We do not claim that the distribution *is* a double exponential. We only claim that the distribution is leptokurtic.

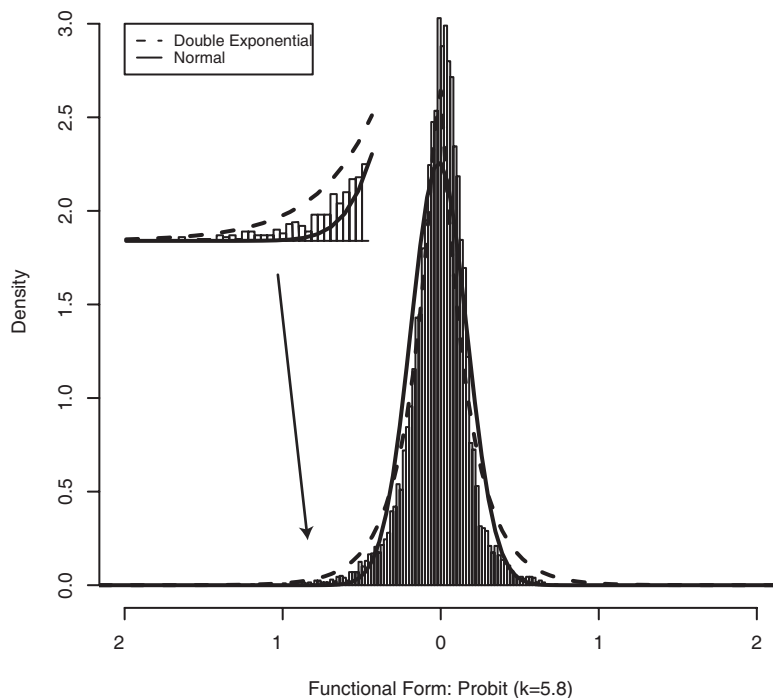


Figure 2: Empirical Distribution of Individual Log-Likelihood Ratios

test to have greater power than the Vuong test. Measuring the power of these tests, however, is not perfectly straightforward. Two issues must be considered: nominal versus exact or natural significance and the definitions of power and size.

### 4.3 A Framework for Comparison

The Vuong statistic is a continuous statistic, while the test statistic for the distribution-free test is discrete. The problem with this comparison is that for any finite number of observations, the exact significance level of the discrete test statistic is unlikely to match the nominal significance level selected for the simulation.<sup>21</sup> Absent identical exact significance levels, power comparisons may be quite misleading (Gibbons and Chakraborti 1992).

---

<sup>21</sup>The distribution-free test has at most  $n + 2$  available  $\alpha$ -levels. These probabilities are known as exact or natural significance levels. Nominal significance levels are those chosen by the researcher in advance of performing a test.

One way to get around this problem is to employ a randomized decision rule (Lehmann 1986). Let a test statistic  $\tau$  be in the rejection region with probability 1 if  $\tau \geq c_2$  and with probability  $\rho$  if  $c_1 \leq \tau \leq c_2$ . The nominal significance level can therefore be achieved, even with a small- $n$  discrete test statistic, using the following rule:

$$\Pr(\tau \geq c_1 | H_0) + \rho \cdot \Pr(c_1 \leq \tau \leq c_2 | H_0) = \alpha, \quad (34)$$

where  $c_1 < c_2$ .

As DeGroot (1989) points out, however, it seems odd for a researcher to decide which hypothesis to accept by tossing a coin or using some other method of randomization. In place of a randomized procedure, we chose critical values for the Vuong test such that the significance level of the Vuong would match the natural significance level of the distribution-free test. Thus, for a sample size of 200, we chose 111 as the critical value for the distribution-free test, which gives a natural  $\alpha$ -level of 0.0518. We then chose a critical value of 1.627 for the Vuong test, which gives the same  $\alpha$ -level. The power levels we report, therefore, are for equivalent exact or natural significance levels.

The concept of power also requires discussion when considering model selection tests. Power is commonly defined as the probability of rejecting a false null hypothesis,

$$\text{Power} = 1 - \beta(\theta) = 1 - \Pr(\tau \notin R | \theta \in \Omega - \omega). \quad (35)$$

where  $R$  is the rejection region, and the hypotheses are  $H_0 : \theta \in \omega$  versus  $H_1 : \theta \in \Omega - \omega$ . For both the Vuong test and distribution-free test, however, we are interested in the probability of rejecting a false null in a particular direction. It therefore makes sense to substitute the probability of making a correct decision for the probability of simply rejecting a false null.

The concept of size raises issues as well. The design of the simulation does not include an experiment where the null hypothesis is true; size therefore cannot be calculated. Given the sample sizes with which we are working, we cannot consistently generate two models that are exactly the same distance away from the DGP. This drawback of the experiment is far from fatal. If the rival models are truly nonnested (see Pesaran (1987) for a technical definition based on the KLIC), the likelihood that they are equally close to the true model is vanishingly small. The probability of rejecting a true null is simply

not an issue for the practical researcher. Rejecting a false null in the wrong direction, however, is an issue to which the practical researcher must pay attention. It again seems reasonable to substitute this practical probability, the probability of making an incorrect decision, for the probability of rejecting a true null.<sup>22</sup>

#### 4.4 Simulation Results 2: Covariates

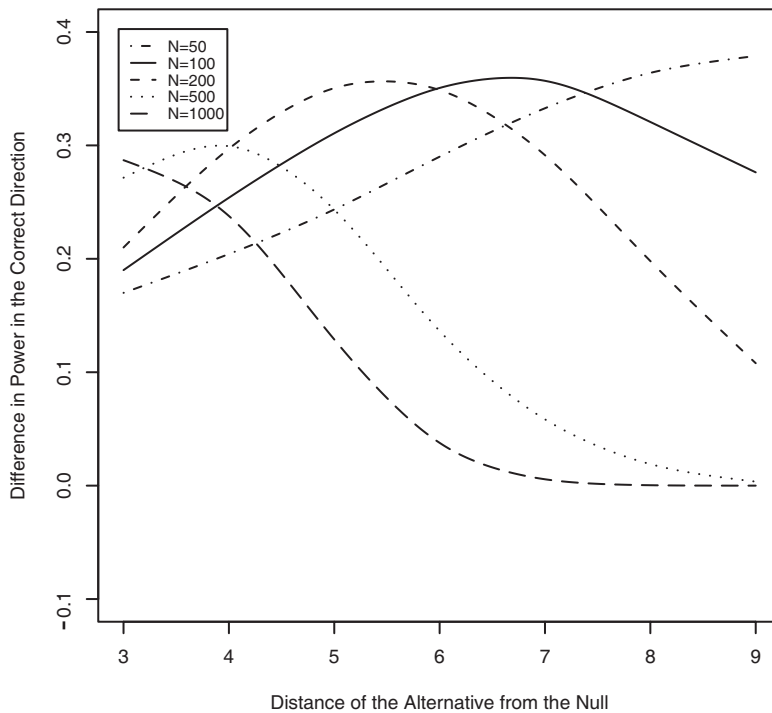


Figure 3: Difference in the Power Functions of the Tests,  $\sigma = 1.0$ ,  $\alpha \approx 0.05$

Figure 3 shows the difference (distribution-free minus Vuong) in the “power” functions of the two tests for an error standard deviation of 1.0 and a sig-

<sup>22</sup>We would, however, like to have the power analysis be finer-grained in the sense of having the alternative hypothesis approach the null more closely than a canonical correlation of 0.3 versus 0.2. With sample sizes of 50 and 100, we simply cannot produce competing models with canonical correlations that are accurate to the second decimal place. Short of a new procedure for drawing random variates from a multivariate normal distribution, jumps of 0.1 between canonical correlations are as fine-grained as we can achieve.

nificance level of approximately 0.05.<sup>23</sup> What is immediately obvious is that the power of the distribution-free test is as great or greater than the power of the Vuong test across all alternatives and sample sizes. For a sample size of 200 and a distance of 5, for instance, the Vuong test chose the correct model in only 17.7% of the replications while the distribution-free test chose the correct model in 53.8% of the replications (a point which appears on the graph as  $53.8\% - 17.7\% = 36.1\%$ ). In general, the power differential between the tests decreases as the sample size increases, reflecting the consistency of both tests.

That being said, the power differential between the tests is a complex interaction between the size of the sample and the distance of the alternative from the null. For a sample size of 50, the largest differential is found at alternatives far from the null (distances of 8 and 9). For sample sizes of 100 and 200, the largest differential is found for alternatives midway from the null (distances of 5 to 7). For sample sizes of 500 and 1000, the largest differential is found at alternatives closest to the null (distances of 3 to 4). This last differential does not disappear quickly as the sample size continues to grow. Re-running the experiment for a sample size of 5000 and a distance of 3, the distribution-free test chose the correct model in 81% of the replications while the Vuong chose the correct model in only 59% of the replications.

The greater power of the distribution-free test does not come without a price. Figure 4 shows the difference (distribution-free minus Vuong, on the same scale as the previous graph) in the “size” of the two tests for an error standard deviation of 1.0 and a significance level of approximately 0.05. It is immediately obvious that the distribution-free test chose the wrong model more often than the very conservative Vuong test. In absolute terms, however, neither test chose the wrong model often, and the probability of either test choosing the wrong model decreased as the sample size increased. At its worst ( $N = 50$ , distance=3), the Vuong test chose the incorrect model in only 3.0% of the replications while the distribution-free test, at its worst ( $N = 100$ , distance=3), chose the wrong model in 12.5% of the replications. Figure 4, then, essentially shows the percentage of replications in which the

---

<sup>23</sup>The results of the Monte Carlo experiment are presented graphically to avoid a large number of tables from which it would be difficult to garner relevant information. A downside to graphical presentations is that the inclusion of standard errors would render the graphs incomprehensible. All standard errors, therefore, are available from the authors upon request.

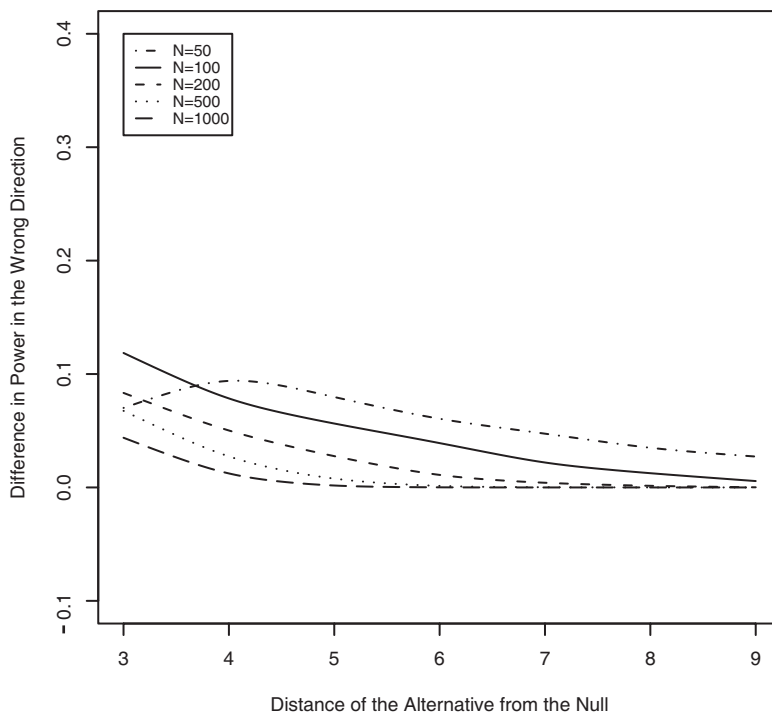


Figure 4: Difference in the Probability of Choosing the Wrong Model,  $\sigma = 1.0$ ,  $\alpha \approx 0.05$

distribution-free test chose the wrong model.<sup>24</sup>

How are we to interpret these results? The Vuong test is far more conservative than the distribution-free test and therefore does a better job of protecting against rejecting the null in the wrong direction. On the other hand, Lehmann (1986) notes that there is little point in carrying out a test that has only a small chance of detecting a false null. Interpreting our results requires finding a balance between these two errors. In general, the best test is the one that minimizes some linear combination of the probability of rejecting a false null in the wrong direction and the probability of failing to reject a false null.

Let  $\alpha(\delta)$  be the probability of failing to reject a false null, and  $\beta(\delta)$  be the probability of rejecting a false null in the wrong direction. Given

---

<sup>24</sup>If we change the error standard deviation from 1.0 to 2.0, both tests perform worse in an absolute sense, although the relative advantage of the distribution-free test increases. Otherwise, the conclusions drawn from the previous two graphs still hold.

positive constants  $a$  and  $b$ , we want to choose the test,  $\delta^*$ , such that the linear combination of the two errors is a minimum,

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta). \quad (36)$$

The question is how much weight to give each of these errors. In most circumstances, scholars are more concerned about choosing the wrong model than choosing neither model.

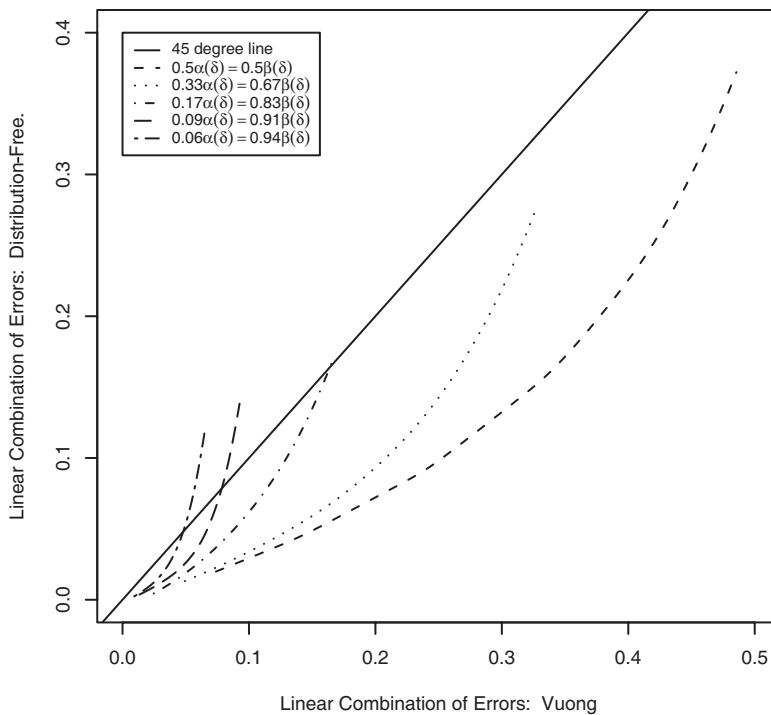


Figure 5: Linear Combinations of Errors,  $n = 200, \sigma = 1.0, \alpha \approx 0.05$

Figure 5 is a representative ( $n = 200, \sigma = 1.0$ ) scatter plot of the linear combination under five weighting schemes. A curve, or any part of a curve, that lies below the 45 degree line indicates that the linear combination of errors for the distribution-free test is less than that of the linear combination of errors for the Vuong test. If, following Pesaran (1974), we assume that two errors are equally important ( $a = b = 0.5$ ), then it follows that the distribution-free test is preferable to the Vuong test. This is the rightmost curve in Figure 5. The distribution-free test remains unambiguously preferred even if the probability of choosing the wrong model is given twice as

much weight ( $a = 0.333, b = 0.667$ ), or even five times the weight ( $a = 0.167, b = 0.833$ ), of the probability of choosing neither model.

If the probability of choosing the wrong model is given ten ( $a = 0.09, b = 0.91$ ) or fifteen ( $a = 0.0625, b = 0.9375$ ) times the weight of the probability of choosing neither model, then the linear combination turns in favor of the Vuong test for the two alternatives nearest the null (distances 3 and 4). The reason is simple. For a sample size of 200 and alternatives near the null, the distribution-free test chose the wrong model more often than the Vuong, as shown in Figure 4. If we then put extraordinary weight on the probability of choosing the wrong model, the linear combination turns in favor of the Vuong. Of course, as the sample size increases past 1000, these leftmost curves approach and then pass the 45 degree line. Therefore, even if we consider choosing the wrong model to be the more serious error, under most circumstances the distribution-free test outperforms the Vuong.

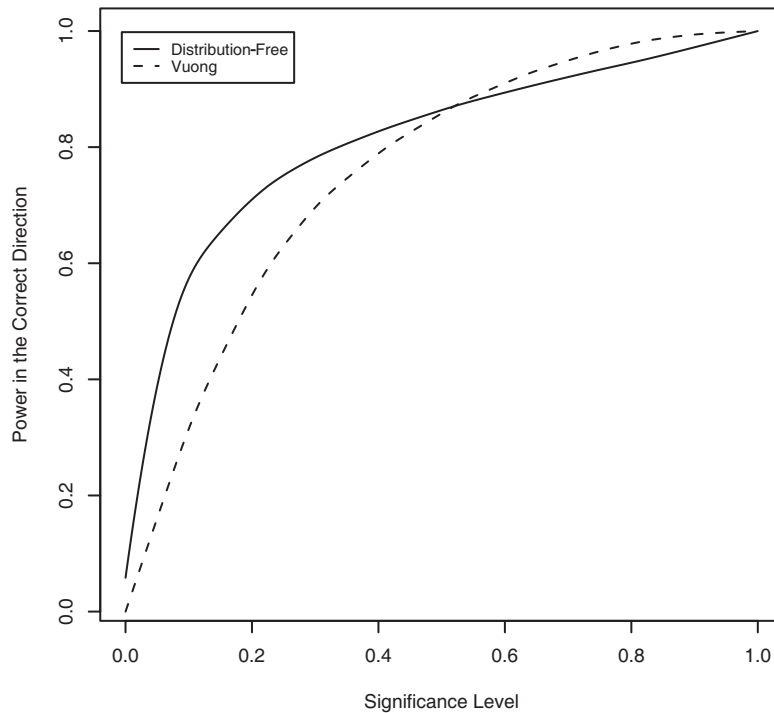


Figure 6: Probability of Choosing the Correct Model by Significance,  $n = 200, \sigma = 1.0, \text{ alternative} = 0.5$

Finally, as power is affected by significance level, we need to assess the

probability of both tests choosing the correct model under different levels of  $\alpha$  (Davidson and MacKinnon 1993). Figure 6 is a representative graph showing the effect of significance level for an alternative that is mid-distance from the null. The distribution-free test has greater power for all reasonable values of  $\alpha$ .

### 4.5 Simulation Results 3: BIC

The final question to be addressed in this paper is how well the distribution-free test performs against model selection criteria, such as the Bayesian information criterion.<sup>25</sup> As noted above, model selection tests are to be preferred over model selection criteria because tests provide a measure of uncertainty regarding model discrimination. That being said, model selection criteria do see significant use in political science, and some comparative results are of interest.

The criterion for model  $i$  is defined as,

$$\text{BIC}_i = -2 \ln f(Y_t | \hat{\beta}) + k_i \ln(n), \quad (37)$$

where  $k_i$  is the number of unknown parameters in model  $i$  (the dimension of  $\beta$ ) (Schwarz 1978). The procedure is quite simple: choose the model for which  $\text{BIC}_i$  is a minimum.

The experiment detailed at the start of this section also produced a decision based on the minimum BIC for each replication. Table 2 reports the difference in the percentage of replications where the BIC and the distribution-free test chose the wrong model (the model further away from the DGP). Across all samples and alternatives, the BIC chooses the wrong model as often or more often than the distribution-free test (and therefore the Vuong test). The BIC chose the wrong test 46% of the time at its worst ( $N = 50$ , distance=3), as compared to 12.5% for the distribution-free test ( $N = 100$ , distance=3). At the same time, however, the BIC chose the correct model more often than the distribution-free test.

As in the last section, we need to look at a linear combination of errors to compare more accurately these approaches. The BIC can make one error, choosing the wrong model, while the model selection tests can make two errors, choosing the wrong model and choosing neither model. Figure 7 is a

---

<sup>25</sup>We focus on BIC as AIC has been shown to be inconsistent (Schwarz 1978), and it was derived in the context of nested models (Ripley 2002).

Distance	Sample Size				
	50	100	200	500	1000
3	0.360	0.295	0.286	0.203	0.149
4	0.291	0.253	0.189	0.085	0.035
5	0.262	0.188	0.103	0.029	0.005
6	0.223	0.125	0.055	0.006	0.000
7	0.167	0.082	0.024	0.001	0.000
8	0.129	0.049	0.007	0.000	0.000
9	0.084	0.025	0.002	0.000	0.000

Table 2: Difference in the Probability of the BIC Criterion and the Distribution-Free Test (BIC minus D.F.) Choosing the Wrong Model,  $\sigma = 1.0$ .

representative scatter plot ( $n = 200$ ,  $\sigma = 1.0$ ) of the linear combination under the same five weighting schemes as in Figure 5. If we assign equal weight to the errors, or twice the weight to choosing the wrong model, the linear combination is in favor of the BIC. This result occurs for two reasons. First, by assigning a weight of 0.5 to the probability that the BIC chose the wrong model, we cut the probability of the BIC making an error, any error, in half. Second, the distribution-free test chose neither model a significant portion of the time. The same reasoning holds if we give twice as much weight to the probability of choosing the wrong model. In this case, we cut the probability of the BIC making an error by a third, and the distribution-free test continues to choose neither model in a significant number of replications. However, if we increase the weight put on the probability of choosing the wrong model to five times that of choosing neither model, the linear combination turns in favor of the distribution-free test.

The results indicate that if we are willing to forgo the probabilistic framework of the model selection tests, we might choose the BIC provided that we are relatively unconcerned about the possibility of choosing the wrong model. If, however, we would like to make a probabilistic statement about model choice and would rather choose neither model over choosing the wrong model, then the distribution-free test is the better choice.

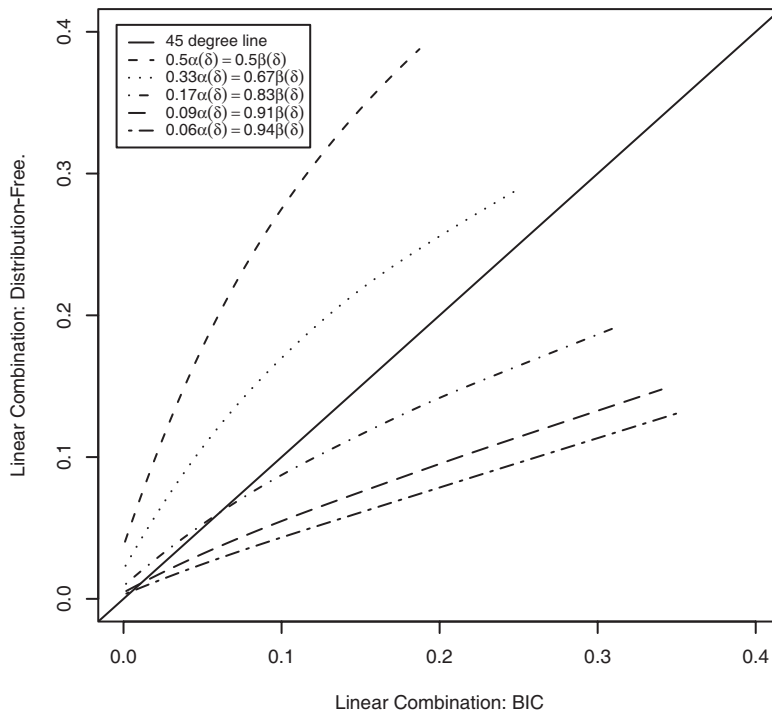


Figure 7: Linear Combinations of Errors,  $n = 200, \sigma = 1.0, \alpha \approx 0.05$

## 5 Conclusion

In this paper, we more fully developed the properties of the distribution-free test for nonnested model discrimination introduced by Clarke (2003). The test meets both technical criteria (consistency, unbiasedness) and practical criteria (ease of calculation and interpretation). We demonstrated that the test is asymptotically more efficient than the well-known Vuong test for highly leptokurtic distributions such as the double exponential. We then demonstrated, through a Monte Carlo experiment, that the distribution of individual log-likelihood ratios is actually highly leptokurtic. Finally, we used the same Monte Carlo to measure the performance of the distribution-free test against the Vuong test and BIC. The results of the first comparison indicate that the distribution-free test outperforms the Vuong test under most circumstances. The results of the second comparison are mixed and depend heavily on the amount of concern one has for rejecting a false null in wrong direction.

As Vuong (1989) noted, “much work remains to be done.” First, although our results regarding the distribution of the individual log-likelihood ratios are highly suggestive, a full characterization of this distribution would prove invaluable. Second, the Monte Carlo experiment reported here should be extended to models that are nonnested solely in terms of their functional forms. Doing so, however, requires finding a family of functional forms such that we can calibrate the distance of various alternatives from the null, much as we did here with canonical correlations. Third, realizing that no particular correction for the number of parameters in the rival models can be justified, it would be enlightening to compare the adjustments suggested here and by Vuong to bootstrapped version of both tests. The results could help indicate which adjustments work best under what conditions. Finally, the extension of the distribution-free test to situations in which there are many competing models would be a greatly anticipated development.

As the number of models and specifications available to researchers continues to increase, nonnested hypothesis testing is likely to become the rule, rather than the exception. A simple, easily interpretable test that provides good power under difficult conditions is surely welcome.

## A Properties of the Distribution-Free Test

Two very intuitive and desirable properties that any useful hypothesis test should possess are consistency and unbiasedness. In the following two sections, these properties are proved for the distribution-free test.

### A.1 Consistency

A consistent test is one that rejects a false null hypothesis with probability one asymptotically.

**Definition A.1** (*Fraser 1957*) *If  $T_{m,n}$  denotes a sequence of size- $\alpha$  tests of  $H_0 : \theta \in \omega$  versus  $H_1 : \theta \in \Omega - \omega$ , then the sequence is said to be consistent for  $\zeta \subset \Omega - \omega$  if*

$$\lim_{n \rightarrow \infty} \mathcal{P}_{T_{m,n}}(\theta) = 1 \quad (38)$$

for  $\theta \in \zeta$ .

To prove consistency, we can make use of the following theorem.

**Theorem A.1** (*Lehmann 1951*) *Let  $T_n$  denote a sequence of test statistics for an  $\alpha$ -level test of  $H_0 : \theta \in \omega$  versus  $H_1 : \theta \in \Omega - \omega$ , such that the test based on  $T_n$  rejects  $H_0$  if  $T_n \geq c_n$  or  $T_n \leq c'_n$ . Suppose there exists a function  $g(\theta)$  such that  $T_n$  converges in probability to  $g(\theta)$  for every  $\theta \in \Omega$ . If, in addition,*

$$g(\theta) = g_0 \quad \forall \theta \in \omega, \quad (39)$$

$$g(\theta) \neq g_0 \quad \forall \theta \in \Omega - \omega, \quad (40)$$

and

$$\lim_{n \rightarrow \infty} c_n \leq g_0, \quad (41)$$

$$\lim_{n \rightarrow \infty} c'_n \geq g_0 \quad (42)$$

then  $T_n$  is a consistent sequence of tests for all alternatives in  $H_1 : \theta \in \Omega - \omega$ .

Let  $\theta$  be the median, and let  $g(\theta) = \Pr(D_i > \theta)$  where  $D_i = \ln \frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}$ .  $g(\theta) = \frac{1}{2}$  for  $\theta \in \omega$ , and  $g(\theta) \neq \frac{1}{2}$  for  $\theta \in \Omega - \omega$ . Now, write the test statistic as,

$$T_n = \frac{B}{n}, \quad (43)$$

where  $B = \sum_{i=1}^n I_{(\theta, +\infty)}(d_i)$ . We show that  $T_n$  converges in probability to  $g(\theta)$  by showing convergence in quadratic mean. The expected value and variance of  $T_n$  are,

$$\begin{aligned} E \left[ \frac{B}{n} \right] &= \frac{1}{n} \sum_{i=1}^n E[I_{(\theta, +\infty)}(d_i)] \\ &= \frac{1}{n} \sum_{i=1}^n [0 * g(\theta) + 1 * g(\theta)] \\ &= g(\theta), \end{aligned} \quad (44)$$

$$\begin{aligned} V \left[ \frac{B}{n} \right] &= \frac{1}{n^2} \sum_{i=1}^n [I_{(\theta, +\infty)}(d_i)] \\ &= \frac{1}{n^2} \sum_{i=1}^n [0^2 * g(\theta) + 1^2 * g(\theta) - g(\theta)^2] \\ &= \frac{g(\theta)[1 - g(\theta)]}{n}. \end{aligned} \quad (45)$$

The variance tends to zero as  $n \rightarrow \infty$ , so we have shown that  $T_n \xrightarrow{p} g(\theta)$ .

As consistency is a large-sample property, we can consider the large-sample approximation of the distribution-free test. Under the null hypothesis, the expected value of  $T_n$  (from above) is  $\frac{1}{2}$ , and the variance is  $\frac{1}{4n}$ . For large  $n$ , the test therefore rejects the null when,

$$\frac{\left| \frac{B}{n} - \frac{1}{2} \right|}{\sqrt{\frac{1}{4n}}} \geq z_{\frac{\alpha}{2}}, \quad (46)$$

where  $z_{\frac{\alpha}{2}}$  is the upper  $100(\alpha/2)$ th percentile of the standard normal. Rearranging, we see that the test rejects when,

$$\left| \frac{B}{n} \right| \geq c_n = \frac{1}{2} + z_{\frac{\alpha}{2}} \sqrt{\frac{1}{4n}}. \quad (47)$$

As  $n \rightarrow \infty$ ,  $c_n \rightarrow \frac{1}{2} \leq g_0 = \frac{1}{2}$ . We have met the conditions of Lehmann's theorem, thus we can state that the test is consistent for all alternatives in  $H_1 : \theta \in \Omega - \omega$ . (Similar proofs can be given for one-tailed tests.)

## A.2 Unbiasedness

An unbiased test is one where the probability of rejection under the null hypothesis is never larger than the probability of rejection under the alternative.

**Definition A.2** (*Lehmann 1986*) A test  $T_{m,n}$  of  $H_0 : \theta \in \omega$  versus  $H_1 : \theta \in \Omega - \omega$  with power function  $\mathcal{P}_{T_{m,n}}(\theta)$  is unbiased of size- $\alpha$  if

$$\mathcal{P}_{T_{m,n}}(\theta) \leq \alpha \quad \forall \theta \in \omega \quad (48)$$

$$\mathcal{P}_{T_{m,n}}(\theta) \geq \alpha \quad \forall \theta \in \Omega - \omega. \quad (49)$$

The distribution-free test is,

$$H_0 : \text{median}_0(D_i) = \theta_0 \text{ versus} \quad (50)$$

$$H_1 : \text{median}_0(D_i) < (>) \theta_0, \quad (51)$$

where  $D_i = \ln \frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}$ . Let  $D_1, \dots, D_n$  be i.i.d.  $F(d - \theta)$ , where  $\theta$  is the median of the underlying distribution. We prove unbiasedness by noting that the distribution-free test reaches its natural significance level for every distribution in  $F(d - \theta)$ , and that its power function is monotonic. We prove the later point using the following theorem.

**Theorem A.2** (*Randles and Wolfe 1979*) Suppose that for testing  $H_0$  versus  $H_1$  we reject  $H_0$  for large (small) values of a test statistic  $T(X_1, \dots, X_n)$  that satisfies

$$T(x_1 + k, \dots, x_n + k) \geq (\leq) T(x_1, \dots, x_n) \quad (52)$$

for every  $k \geq 0$  and  $(x_1, \dots, x_n)$ . Then the test has a monotone power function in  $\theta$  for the one-sample location problem; that is,

$$\mathcal{P}_T(\theta, F) \leq \mathcal{P}_T(\theta', F) \text{ for } \theta \leq \theta', \quad (53)$$

and any continuous distribution with c.d.f.  $F(\cdot)$ .

The distribution-free test rejects for large (small) values of

$$B(D_1, \dots, D_n) = \sum_{i=1}^n I_{(\theta_0, +\infty)}(d_i) \quad (54)$$

where  $I$  is the indicator function. When  $k \geq 0$ ,

$$\begin{aligned} B(d_1 + k, \dots, d_n + k) &= \sum_{i=1}^n I_{(\theta_0, +\infty)}(d_i + k) \\ &= \sum_{i=1}^n I_{(\theta_0 - k, +\infty)}(d_i) \\ &\geq B(d_1, \dots, d_n) \end{aligned} \quad (55)$$

The test then has a monotone power function in  $\theta$ , and therefore the distribution-free test is an unbiased of  $H_0 : \text{median}_0(D_i) = \theta_0$  against  $H_1 : \text{median}_0(D_i) > (<) \theta_0$ .

## References

- Akaike, H. 1973. "Information Theory and an Extension of the Likelihood Ratio Principle." In *Second International Symposium of Information Theory*, eds. B.N. Petrov and F. Csaki. Minnesota Studies in the Philosophy of Science, Budapest: Akademinai Kiado.
- Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.
- Amemiya, Takeshi. 1980. "Selection of Regressors." *International Economic Review* 21:331–347.
- Balanda, Kevin P., and H. L. MacGillivray. 1988. "Kurtosis: A Critical Review." *The American Statistician* 42:111–119.
- Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.
- Bradley, James V. 1968. *Distribution-Free Statistical Tests*. New Jersey: Prentice-Hall.
- Clarke, Kevin A. 2001. "Testing Nonnested Models of International Relations: Reevaluating Realism." *American Journal of Political Science* 45:724–744.
- Clarke, Kevin A. 2003. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47:72–93.
- Clarke, Kevin A., and Curt S. Signorino. 2004. "Discriminating Methods: Nonnested Tests for Strategic Choice Models."
- Cox, David R. 1961. "Tests of separate families of hypotheses." *Proceedings of the Fourth Berkeley Symposium* I:105–123.
- Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- DeGroot, Morris H. 1989. *Probability and Statistics*. 2 ed. Reading, MA: Addison-Wesley.
- Fraser, D. A. S. 1957. *Nonparametric Methods in Statistics*. New York: John Wiley and Sons.
- Gibbons, Jean Dickinson, and Subhabrata Chakraborti. 1992. *Nonparametric Statistical Inference*. 3 ed. New York: Marcel Dekker, Inc.

- Granger, C.W.J., Maxwell L. King, and Halbert White. 1995. "Comments on testing economic theories and the use of model selection criteria." *Journal of Econometrics* 67:173–187.
- Hendry, David F. 1984. "Monte Carlo Experimentation in Econometrics." In *Handbook of Econometrics*, eds. Z. Griliches and M.D. Intriligator, vol. II, chap. 16. Amsterdam: Elsevier Science Publishers BV, 939–976.
- Hodges, J.L., and E.L. Lehmann. 1956. "The Efficiency of Some Non-parametric Competitors of the t-Test." *Annals of Mathematical Statistics* 27:324–335.
- Kaiser, Henry F., and Kern Dickman. 1962. "Sample and Population Score Matrices and Sample Correlation Matrices from an Arbitrary Population Correlation Matrix." *Psychometrika* 27:179–182.
- Kass, Robert, and Adrian Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–795.
- Kullback, Solomon, and R.A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86.
- Lavine, Michael, and Mark J. Schervish. 1999. "Bayes Factors: What They Are and What They Are Not." *The American Statistician* 53:119–122.
- Lehmann, E. L. 1951. "Consistency and Unbiasedness of Certain Non-parametric Tests." *Annals of Mathematical Statistics* 22:165–179.
- Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. 2 ed. New York: John Wiley.
- Noether, Gottfried E. 1955. "On a Theorem of Pitman." *Annals of Mathematical Statistics* 26:64–68.
- Noether, Gottfried E. 1967. *Elements of Nonparametric Statistics*. New York: John Wiley & Sons, Inc.
- Pesaran, M. H. 1974. "On the General Problem of Model Selection." *Review of Economic Studies* 41:153–171.
- Pesaran, M.H. 1987. "Global and partial non-nested hypotheses and asymptotic local power." *Econometric Theory* 3:69–97.

- Pesaran, M.H., and B. Pesaran. 1993. "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Nonnested Models." *Journal of Econometrics* 57:377–392.
- Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research (with discussion)." In *Sociological Methodology 1995*, ed. P.V. Marsden. Cambridge, MA: Blackwells.
- Randles, Ronald H., and Douglas A. Wolfe. 1979. *Introduction to The Theory of Nonparametric Statistics*. New York: John Wiley and Sons.
- Ripley, Brian D. 2002. "Formal versus Informal Methods of Model Choice."
- Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–464.
- Smith, Alastair. 1999. "Testing Theories of Strategic Choice: The Example of Crisis Escalation." *American Journal of Political Science* 43:1254–1283.
- Spanos, Aris. 1999. *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press.
- Vuong, Quang. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica* 57:307–333.
- White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.