Advance Access publication February 24, 2007

Political Analysis (2007) 15:347–363
doi:10.1093/pan/mpm004

A Simple Distribution-Free Test for Nonnested Model Selection

Kevin A. Clarke
Department of Political Science, University of Rochester, Rochester, NY 14627-0146
e-mail: kevin.clarke@rochester.edu

This paper considers a simple distribution-free test for nonnested model selection. The new test is shown to be asymptotically more efficient than the well-known Vuong test when the distribution of individual log-likelihood ratios is highly peaked. Monte Carlo results demonstrate that for many applied research situations, this distribution is indeed highly peaked. The simulation further demonstrates that the proposed test has greater power than the Vuong test under these conditions. The substantive application addresses the effect of domestic political institutions on foreign policy decision making. Do domestic institutions have effects because they hold political leaders accountable, or do they simply promote political norms that shape elite bargaining behavior? The results indicate that the latter model has greater explanatory power.

1 Introduction

How do domestic political institutions affect foreign policy decision making? Huth and Allee (2002) compare three models corresponding to three different causal mechanisms that link domestic institutions to foreign policy decisions. Deciding which model best fits the data requires choosing between nonnested models, which is a problem too rarely discussed in quantitative political science. Although Bayesian statisticians have made significant inroads on the problem (Schwarz 1978; Carlin and Chibb 1995; Laud and Ibrahim 1995; Albert 1996; Berger and Pericchi 1996; Brown, Vannucci, and Fearn 1998; George and Foster 2000; Fernandez, Ley, and Steel 2001), the fact remains that many statisticians, never mind political scientists, are not Bayesians (Efron 1986).[1] Unfortunately, non-Bayesian approaches to the problem of nonnested model testing have not received the same attention.

A notable exception is Vuong (1989), who introduces a directional and symmetric test for choosing between nonnested models. In a similar vein, this paper considers a directional and symmetric distribution-free test for nonnested model selection introduced by Clarke (2003). Although both tests are consistent and unbiased, we show that the

Authors' note: This work was supported by National Science Foundation Grant SES-0213771. I thank Paul K. Huth and Todd L. Allee for graciously sharing their data and code. I also thank Bear Braumoeller, Curtis Signorino, Tasos Kalandrakis, participants in the NorthEast Methodology Program, New York University, 2003, and the reviewers for their comments. Errors remain my own. Supplementary materials are available on the *Political Analysis* Web site.

[1] An additional problem is that the implementation of many, but not all, Bayesian methods for model selection is far from straightforward; see Chipman, George, and McCulloch (2001).

distribution-free test is asymptotically more efficient than the Vuong test when the distribution of the underlying data is highly peaked (relative to the Normal distribution). This result is important because in many applied research situations where the rival models are characterized by nonnested design matrices, the underlying distribution is highly peaked. The distribution-free test significantly outperforms the Vuong test in these situations.

The substantive application demonstrates the point. Whereas the Vuong test cannot discriminate between Huth and Allee's (2002) political accountability model and political norms model, the distribution-free test does distinguish between them. The results suggest that the political norms model has greater explanatory power than previously thought and merits further investigation.

The article is organized as follows. Section 2 reviews the Vuong test and the distribution-free test. Section 3 compares the asymptotic relative efficiency of the tests for Normal and highly peaked distributions. Section 4 describes the setup of the simulation study and discusses the results. Section 5 introduces the substantive debate and provides the results.

2 The Vuong and Distribution-Free Tests

The Vuong test (Vuong 1989) and the distribution-free test (Clarke 2003) are based on the Kullback-Leibler information criteria.[2] Consider a model conditioned on some covariates, $\mathbf{F}_\beta = f(Y_i|X_i;\beta)$. As defined by Vuong, the Kullback-Leibler distance is KLIC $\equiv E_0[\ln h_0(Y_i|X_i)] - E_0[\ln f(Y_i|X_i;\beta_*)]$, where $h_0(\cdot|\cdot)$ is the true conditional density of Y_i given X_i, E_0 is the expectation under the true model, and β_* are the pseudo-true values of β (see White 1982). The model that minimizes the KLIC 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 $E_0[\ln f(Y_i|X_i;\beta_*)]$.

2.1 *The Vuong Test*

Consider two models, $\mathbf{F}_\beta = f(Y_i|X_i;\beta)$ and $\mathbf{G}_\gamma = g(Y_i|Z_i;\gamma)$. The null hypothesis of the test is

$$H_0 : E_0\left[\ln\frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}\right] = 0, \tag{1}$$

which indicates that two rival models are equally close to the true specification. Vuong proves under general conditions that the expected value given in the null hypothesis can be consistently estimated by $(1/n)$ times the likelihood ratio statistic,

$$\frac{1}{n}\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} E_0\left[\ln\frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}\right], \tag{2}$$

where $\hat{\beta}_n$ and $\hat{\gamma}_n$ are the maximum likelihood estimators of β_* and γ_*. The resulting likelihood ratio statistic is asymptotically normally distributed,[3] and the actual test is therefore

$$\text{under } H_0 : \frac{\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0,1), \tag{3}$$

[2]The Kullback-Leibler information criteria (Kullback and Leibler 1951) is a measure of closeness that has been extensively used in the development of model discrimination procedures due to its analytic tractability and useful properties. See Pesaran (1987) and Clarke (2001) for further details.

[3]See Cameron and Trivedi (2005, 146) for the necessary quasi-maximum likelihood result.

where the numerator is the difference in the summed log-likelihoods for the two models, $\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv L_n^f(\hat{\beta}_n) - L_n^g(\hat{\gamma}_n)$, and $\hat{\omega}_n$ is the estimated SD calculated in the usual manner,

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f(Y_i|X_i;\hat{\beta}_n)}{g(Y_i|Z_i;\hat{\gamma}_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(Y_i|X_i;\hat{\beta}_n)}{g(Y_i|Z_i;\hat{\gamma}_n)}\right]^2. \tag{4}$$

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

$$\tilde{\mathrm{LR}}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv \mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n) - \left[\left(\frac{p}{2}\right)\ln n - \left(\frac{q}{2}\right)\ln n\right], \tag{5}$$

where p and q are the number of estimated coefficients in models $\mathbf{F}_\beta$ and $\mathbf{G}_\gamma$, respectively.

2.2 *The Distribution-Free Test*

Clarke's (2003) distribution-free alternative applies a modified paired sign test to the differences in the individual log-likelihoods from two nonnested models.[4] Using Vuong's notation, the null hypothesis of the distribution-free test is

$$H_0 : Pr_0\left[\ln\frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)} > 0\right] = 0.5. \tag{6}$$

Equation (6) states that under the null hypothesis, the log-likelihood ratios should be evenly distributed around zero. Thus, half the log-likelihood ratios should be greater than zero and half less than zero. The difference between equation (6) and equation (1) is that the expectation in equation (1) is replaced with the median in equation (6).[5]

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

$$B = \sum_{i=1}^{n} I_{(0,+\infty)}(d_i), \tag{7}$$

where I is the indicator function. Equation (7) is the number of positive differences, and it is distributed Binomial with parameters n and $\theta = 0.5$.

If model $\mathbf{F}_\beta$ is "better" than model $\mathbf{G}_\gamma$, 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_α is chosen to be the smallest integer such that $\sum_{c=c_\alpha}^{n}\binom{n}{c}0.5^n \leq \alpha$. For a lower tail test, the inequality is reversed, and the sum goes from $c = 0$ to $c = c_\alpha$.

This test, like the Vuong test, is sensitive to the dimensionality of the competing models. As we are working with the individual log-likelihood ratios, we cannot apply

[4] For those used to thinking of the paired sign test in terms of treatments, it is straightforward to view rival model specifications as treatments. See Clarke (2007) for further detail.

[5] The two assumptions of the test are unsurprising and quite general: the differences, $\ln [f(Y_i|X_i;\beta_*)/g(Y_i|Z_i;\gamma_*)]$, are mutually independent, and each $\ln [f(Y_i|X_i;\beta_*)/g(Y_i|Z_i;\gamma_*)]$ comes from a continuous population (not necessarily the same) that has a common median θ. Proofs of the consistency and unbiasedness of the distribution-free test are in the supplementary materials available on the *Political Analysis* Web site.

the same 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 $\mathbf{F}_\beta$ by a factor of $[(p/2n) \ln n]$ and the individual log-likelihoods for model $\mathbf{G}_\gamma$ by a factor of $[(q/2n) \ln n]$.

Although 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, K_n, divided by the square root of n has a stochastic order of 1, $n^{-1/2}K_n(\mathbf{F}_\beta, \mathbf{G}_\gamma) = o_p(1)$, the adjusted statistic has the same asymptotic properties of the unadjusted statistic. This argument 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 the supplementary materials available on the *Political Analysis* Web site), we see that the asymptotic properties of the distribution-free test are also unaffected by the correction.

3 Comparing the Tests

The distribution-free test, unlike the Vuong test, considers only whether an individual log-likelihood ratio is greater or less than zero, not the degree to which the ratio is greater or less than zero. Thus, the distribution-free test might appear to sacrifice too much of the available information. This criticism is a common, but often mistaken, argument against the use of distribution-free tests, which can be significantly more efficient than their normal theory competitors when the underlying populations are not normal (Hollander and Wolfe 1999, 1). An objective criterion for balancing the trade-offs between these tests is Pitman efficiency, or asymptotic relative efficiency, credited to Pitman in an unpublished paper with subsequent generalizations by Noether (1955) and Hodges and Lehmann (1956).

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^*,

$$\text{A.R.E.}(T, T^*) = \lim_{n \to \infty} \frac{\text{eff}(T_n)}{\text{eff}(T_n^*)}, \tag{8}$$

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

$$\text{eff}(T_n) = \frac{[\mathrm{d}E(T_n)/\mathrm{d}\theta]^2|_{\theta = \theta_0}}{\text{Var}(T_n)|_{\theta = \theta_0}}. \tag{9}$$

Proof of the equivalence between the limiting efficacy ratio and asymptotic relative efficiency, along with the regularity conditions, is given by Gibbons and Chakraborti (1992, chap. 14). The conditions are quite general, and both tests under consideration meet them.

As noted in Section 2.2, the distribution-free test is based on the paired sign test, the efficacy of which is a standard result given by Noether (1967). For N observations from any population F_D (where $D = X - Y$) with median θ, the efficacy is $\text{eff}(B_n) = 4Nf^2[F^{-1}(0.5)]$. For a Normal distribution, $F_X \sim N(\theta, \sigma^2)$, f_X reduces to $f_X = 1/(\sigma\sqrt{2\pi})$, and the efficacy is $\text{eff}(B_n) = 2N/(\pi\sigma^2)$. For a double exponential (or Laplace) distribution (the reason for choosing this distribution will become clear), f_X reduces to $f_X = 1/(2\lambda)$, and the efficacy is $\text{eff}(B_n) = N/\lambda^2$.

The efficacy of the Vuong test is not difficult to calculate. We can write the test statistic as

$$V_n = \frac{\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{N})\hat{\omega}_n} = \frac{\sqrt{N}\left[\frac{1}{n}\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n)\right]}{\hat{\omega}_n}. \tag{10}$$

Letting $\bar{D} = (1/n)\mathrm{LR}_n(\hat{\beta}_n, \hat{\gamma}_n)$ and $\mu_D = E_0\{\ln[f(Y_i|X_i;\beta_*)/g(Y_i|Z_i;\gamma_*)]\}$, we see that the above is equal to

$$\left[\frac{\sqrt{N}(\bar{D} - \mu_D)}{\omega} + \frac{\sqrt{N}\mu_D}{\omega}\right]\frac{\omega}{\hat{\omega}_n}. \tag{11}$$

Given that $\lim_{N\to\infty}(\hat{\omega}_n/\omega) = 1$ (see Vuong 1989, 351), the expected value and variance of the Vuong test statistic for large N are

$$E[V_n] = \frac{\sqrt{N}\mu_D}{\omega}, \qquad \mathrm{Var}[V_n] = \frac{N\ \mathrm{Var}(\bar{D})}{\omega^2} = 1.$$

The efficacy is therefore $\mathrm{eff}(V_n) = N/\omega^2$. For a Normal distribution, the efficacy of the Vuong is N/σ^2. For the double exponential, the efficacy is $\mathrm{eff}(V) = N/\omega^2 = N/(2\lambda^2)$.

We are now in a position to make some statements regarding the asymptotic relative efficiency of the Vuong test versus the distribution-free test. From the discussion above, it is clear that the difference in the A.R.E. of the two tests depends on the shape of the distribution. In particular, it is the kurtosis value of the distribution that matters. Kurtosis measures whether a symmetric distribution has, relative to the Normal, thicker tails and higher peaks or not (Spanos 1999, 119). The Normal is a mesokurtic distribution and has a kurtosis of 3. The double exponential has thicker tails and is more peaked than the Normal, and thus is a leptokurtic distribution with a kurtosis of 6. The Uniform distribution has no tails and no peak. The Uniform is a platykurtic distribution and has a kurtosis of 9/5. The difference between these distributions can be seen in Fig. 1.

For normally distributed data, the asymptotic relative efficiency of the tests is

$$\mathrm{A.R.E.}(B, V) = \lim_{n\to\infty}\frac{\mathrm{eff}(B_n)}{\mathrm{eff}(V_n)} = \frac{2N/(\pi\sigma^2)}{N/\sigma^2} = \frac{2}{\pi}. \tag{12}$$

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. 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 leptokurtic distributions such as the double exponential. For data that are distributed according to the double exponential, the asymptotic relative efficiency is

$$\mathrm{A.R.E.}(B, V) = \lim_{n\to\infty}\frac{\mathrm{eff}(B_n)}{\mathrm{eff}(V_n)} = \frac{N/\lambda^2}{N/(2\lambda^2)} = 2. \tag{13}$$

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. Under these conditions, we are better off using the distribution-free test, as it provides greater power than the Vuong test.

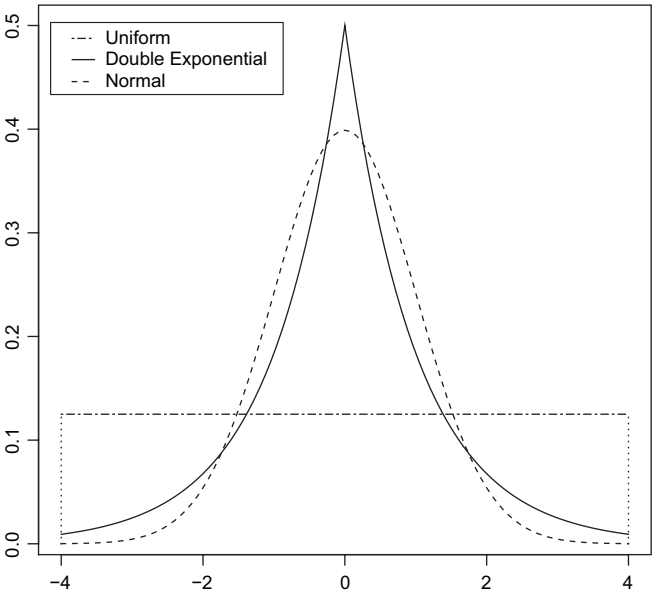

Fig. 1 A platykurtic (the Uniform), a mesokurtic (the Normal), and a leptokurtic (the double exponential) distribution.

4 The Monte Carlo Experiment

Given that we have established that the distribution-free test is asymptotically more efficient for leptokurtic distributions, the Monte Carlo experiment is designed to answer two questions. First, is the distribution of individual log-likelihood ratios for rival models characterized by nonnested design matrices leptokurtic? If so, we should expect the distribution-free test to have greater power than the Vuong test. Second, does the distribution-free test actually have greater power in this situation?

The Monte Carlo experiment comprises three models, one of which is the data-generating process (DGP) and two that vary in distance from each other and the DGP. This setup reflects the fact that substantive researchers rarely have the luxury of choosing between two models, one of which is the DGP. It is far more likely that both models being compared are misspecified in some fashion, and that the true model is unknown. The rival models in the experiment are nonnested in terms of their design matrices, and a binary choice model with a probit link function was chosen for its ubiquity in social science research, as well as its relationship to the Normal distribution.[6] (We might expect that if the distribution of log-likelihood ratios is ever normal, it will be normal for generalized linear models with normal link functions. The simulation was rerun with other functional forms and similar or more extreme results were found.)

4.1 *Setup of the Experiment*

The experiment calls for three models: one true model and two misspecified models. 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 true model along

[6]The objection might be raised that nonnested tests are unnecessary when the models are nonnested solely in term of their design matrices. This is actually not the case, and we address this issue at greater length in Section 5.

with a randomly drawn, normally distributed error term with a SD 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). The canonical correlations control how far the misspecified models are from the true model and each other.

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.

The moving parts of the experiment include the sample size (50, 100, 200, 500, 1000), distance of the alternative from the null (0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), and the error SD (1, 2). Thus, 70 variations on the experiment were performed. Each replication in each variation led to either a rejection or an acceptance of the null hypothesis for both tests. 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, 739). A total of 8000 replications of each variation was run to ensure that the width of the 95% confidence interval around each estimate is approximately 0.01. All coefficients are set to 1, and the six independent variables, as well as the error term, were drawn anew for each replication.

4.2 *Simulation Results 1: Kurtosis*

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.[7]

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 individual log-likelihood ratios are therefore unlikely to be distributed normally.

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. Figure 2 shows one such representative distribution with the Normal and double exponential distributions superimposed. The closer match to the double exponential is clear. (Note that we do not claim that the distribution *is* a double exponential; we only claim that the distribution is leptokurtic. It is a good idea to graph and check the empirical distribution in any substantive application.) Although 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 is more leptokurtic than mesokurtic, we expect the distribution-free test to have greater power than the Vuong test. Measuring the power of these tests, however, is not perfectly straightforward. Two issues

[7]The values were calculated using the kurtosis function of Angelo Mineo's Normalp library in *R*.

Table 1 Mean kurtosis coefficients

	Sample size				
Distance	*50*	*100*	*200*	*500*	*1000*
0.3	5.19	5.44	5.49	5.50	5.49
0.4	5.26	5.52	5.58	5.57	5.59
0.5	5.24	5.49	5.62	5.68	5.73
0.6	5.32	5.62	5.81	5.88	5.90
0.7	5.40	5.70	5.89	6.09	6.14
0.8	5.47	5.87	6.10	6.34	6.38
0.9	5.58	5.98	6.35	6.58	6.70

must be considered: nominal versus exact or natural significance and the definition of power.

4.3 *A Framework for Comparison*

Comparing the two test statistics is complicated by the fact that one is continuous and one is discrete. The exact significance level of the discrete statistic is unlikely to match the nominal significance level selected for the simulation (the distribution-free test has at most $n + 2$ available α levels). Absent identical exact significance levels, power comparisons may be quite misleading (Gibbons and Chakraborti 1992, 21). To solve this problem, 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. (A randomized procedure

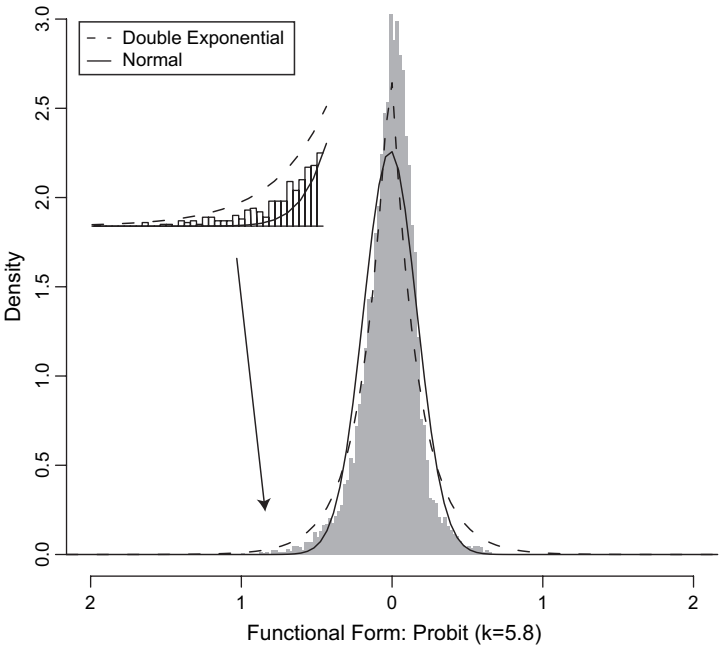

Fig. 2 Empirical distribution of individual log-likelihood ratios.

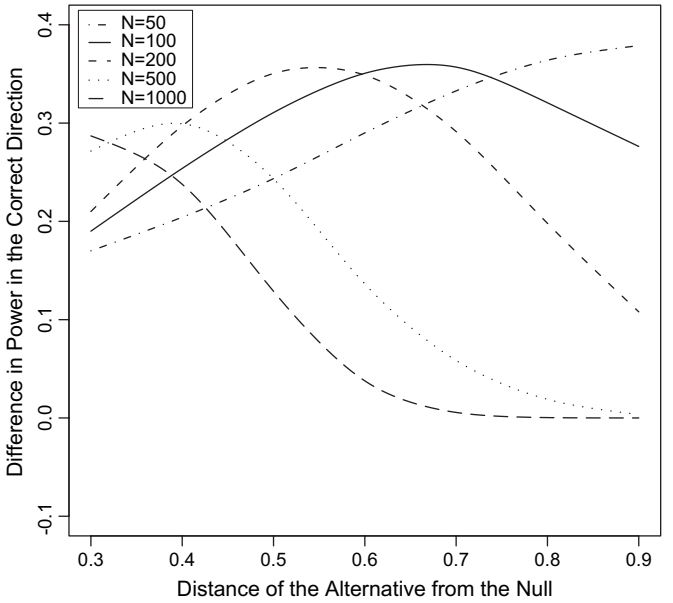

Fig. 3 Difference in the power functions of the tests, $\sigma = 1.0$, $\alpha \approx 0.05$.

would have been another possibility, see Lehmann 1986.) 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. Let Ω be the parameter space and ω be the part of the parameter space that includes the null hypothesis. The hypotheses are therefore H_0: $\theta \in \omega$ versus H_1: $\theta \in \Omega - \omega$, and

$$\text{Power} = 1 - \beta(\theta) = 1 - \Pr(\tau \notin R \,|\, \theta \in \Omega - \omega), \tag{14}$$

where R is the rejection region. For both the Vuong test and the 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 rejecting a false null.

4.4 *Simulation Results 2: Covariates*

Figure 3 shows the difference (distribution-free minus Vuong) in the "power" functions of the two tests for an error SD of 1.0 and a significance level of approximately 0.05. 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 0.5, for instance, the Vuong test chose the correct model in only 17.7% of the replications, whereas 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 nonlinear interaction between the size of the sample and the distance of the alternative from the null. Each line

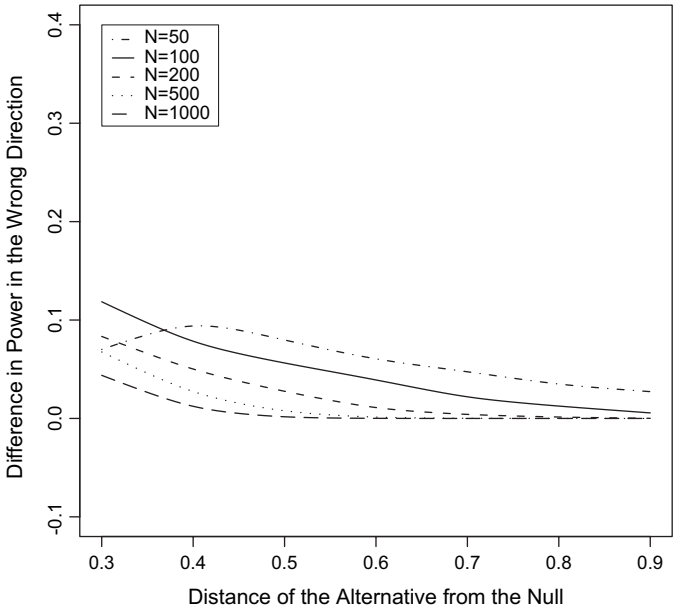

Fig. 4 Difference in the probability of choosing the wrong model, $\sigma = 1.0$, $\alpha \approx 0.05$.

in Fig. 3 is hill shaped, and the location of each hill is determined by the size of the sample. The larger the sample, the farther to the left the hill is centered. The implication is that for a large enough sample, there would be no power difference between the tests. The sample, however, would have to be quite large. When we reran the experiment for a sample size of 5000 and a distance of 0.3, the distribution-free test chose the correct model in 81% of the replications, whereas the Vuong test 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 probability of choosing the wrong model for an error SD 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 $= 0.3$), the Vuong test chose the incorrect model in only 3.0% of the replications, whereas the distribution-free test, at its worst ($N = 100$, distance $= 0.3$), chose the wrong model in 12.5% of the replications.[8]

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. Lehmann (1986), however, 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. One option is choosing a test 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.

[8]If we change the error SD 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.

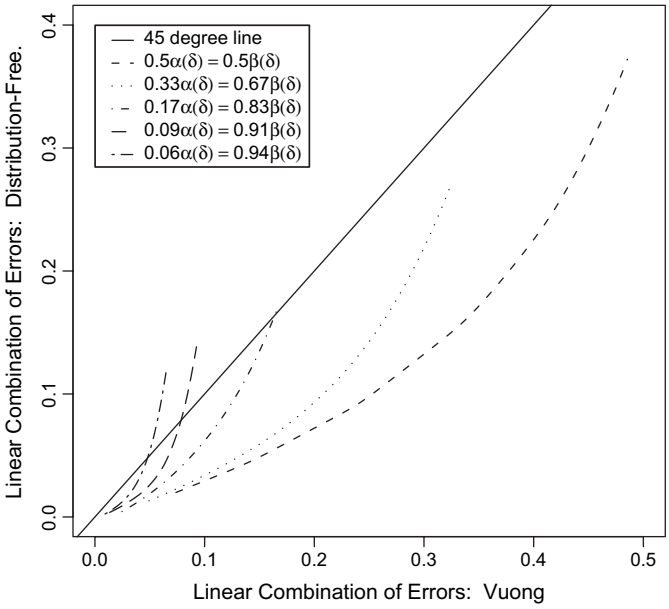

Fig. 5 Linear combinations of errors, $n = 200$, $\sigma = 1.0$, $\alpha \approx 0.05$.

Let $\alpha(\delta)$ be the probability of failing to reject a false null, and let $\beta(\delta)$ be the probability of rejecting a false null in the wrong direction. Given positive constants a and b, we want to choose the test, δ^*, such that the linear combination of the two errors is the smallest, $a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta)$. 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 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 Fig. 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 10 ($a = 0.09$, $b = 0.91$) or 15 ($a = 0.0625$, $b = 0.9375$) 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 0.3 and 0.4). 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 Fig. 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.

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 α (Davidson and

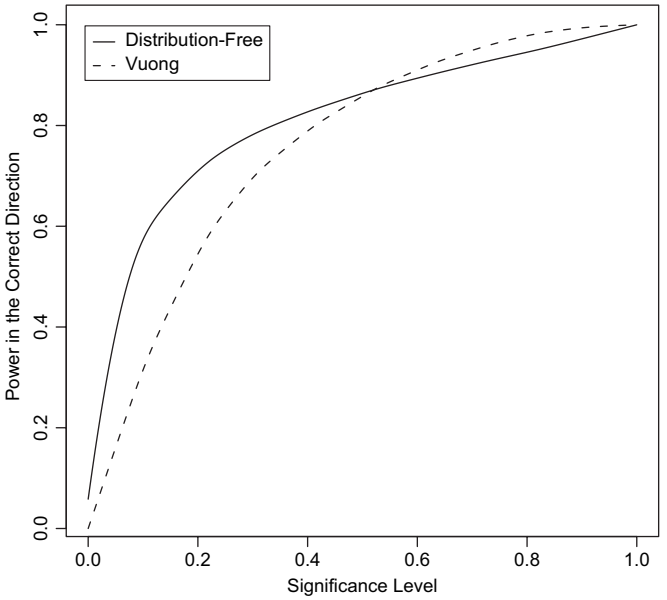

Fig. 6 Probability of choosing the correct model by significance, $n = 200$, $\sigma = 1.0$, alternative $= 0.5$.

MacKinnon 1993, 405). 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 α.

5 Application

How do domestic political institutions affect foreign policy decision making? Huth and Allee (2002) compare three models corresponding to three different causal mechanisms that link domestic institutions to foreign policy decisions.[9] In general, they find that the empirical evidence supports what they call the political accountability model [286]. Although their models are nonnested, Huth and Allee rely on informal methods of model comparison. The application of nonnested tests to their models, however, calls into question some of their findings.

In the political accountability model, "competitive elections, independent legislative powers, and the threat of military coups are the sources of accountability for leadership decisions in foreign policy" (Huth and Allee 2002, 101). The model comprises four key assumptions: the critical goal of incumbent leaders is the retention of their office; political opponents challenge incumbents at strategic junctures; political accountability varies across different domestic political institutions; and the greater the political vulnerability of leaders, the more risk-averse leaders are in their foreign policy.[10]

[9] Although Huth and Allee compare three models, we focus on the two models with the greatest direct support.

[10] Huth and Allee (2002) operationalize the political accountability model using six variables: a measure of how democratic the challenger and target are, whether the dispute is at a stalemate, whether the dispute is part of an enduring rivalry, whether ethnic co-nationals are involved in the dispute, whether the situation is one of high military risk or uncertainty, and the resolve of the target. Each of the last five variables is interacted with democracy in order to understand how the greater accountability of democratic leaders affects the decision to escalate a crisis.

Table 2 The political accountability model, $n = 374$

	Challenger		*Target*	
Variable	*Coefficient*	*SE*	*Coefficient*	*SE*
Challenger level of democracy	−0.004	0.017		
Target level of democracy			0.006	0.017
Democracy × stalemate	−0.030	0.023	−0.008	0.019
Control for stalemate	−0.418	0.192	−0.421	0.160
Democracy × enduring rivalry	0.000	0.018	−0.016	0.014
Control for enduring rivalry	0.106	0.172	0.233	0.159
Democracy × ethnic ties	−0.014	0.018	0.005	0.016
Control for ethnic ties	0.400	0.143	0.053	0.121
Democracy × military risk	0.010	0.022	−0.026	0.016
Control for military risk	−0.141	0.198	0.066	0.229
Target resolve × target democracy	−0.036	0.014		
Target signal of resolve	−0.052	0.118		
Military balance	0.931	0.301	−0.062	0.443
Local balance of forces	0.440	0.144	0.026	0.186
Strategic value	0.559	0.148	0.334	0.142
Common security interests	−0.432	0.184	−0.128	0.175
Target other dispute	0.300	0.169	0.360	0.162
Challenger other dispute	0.351	0.164	0.135	0.164
Constant	−1.916	0.265	−1.237	0.230
ρ	0.956	0.021		
Log-likelihood	−243.063			

In the political norms model, "attention shifts to the principles that shape political elite beliefs about how to bargain and resolve political conflicts," and leaders from democratic and nondemocratic states have "different beliefs about the acceptability of compromising with and coercing political adversaries" (Huth and Allee 2002, 101). The model comprises three key assumptions: norms influence decisions made by political actors in political conflict, domestic political institutions structure political conflict, and the bargaining strategies used by leaders in international disputes are influenced by the norms of bargaining those same leaders use with domestic political opponents.[11]

The models are completed by including a set of straightforward realist variables comprising military balance, local balance of forces advantage, the strategic value of the territory, alliance behavior, and whether or not the target or challenger are involved in another militarized dispute.[12] The models are tested on 374 territorial disputes where the challenger has opted for military pressure over calling for negotiations. The challenger and the target both choose to either escalate the dispute or not (the dependent variable); consequently, the models are estimated by bivariate probit. The results in Tables 2 and 3 replicate the results of Huth and Allee's Tables 9.4 [240] and 9.13 [256], respectively.

A question that arises is how informative the Monte Carlo results are in regard to this application. There are five points of convergence. First, the Monte Carlo experiment comprises two models, neither of which is the DGP. Neither of the two Huth and Allee

[11]Huth and Allee (2002) operationalize the political norms model with four variables: the strength of nonviolent norms in the state, whether the dispute is at a stalemate, nonviolent norms interacted with stalemate, and nonviolent norms interacted with a measure of whether or not the state has a military advantage.

[12]See Huth and Allee (2002) for information on how these variables are operationalized.

Table 3 The political norms model, $n = 374$

	Challenger		*Target*	
Variable	*Coefficient*	*SE*	*Coefficient*	*SE*
Strength of nonviolent norms	−0.009	0.021	−0.020	0.018
Nonviolent norms × military advantage	−0.017	0.018	−0.016	0.014
Nonviolent norms × stalemate	0.005	0.036	0.007	0.043
Control for stalemate	−0.345	0.429	−0.425	0.509
Military balance	1.201	0.377	0.057	0.387
Local balance of forces	0.535	0.152	0.085	0.176
Strategic value	0.485	0.144	0.353	0.144
Common security interests	−0.404	0.190	−0.136	0.176
Target other dispute	0.404	0.167	0.408	0.163
Challenger other dispute	0.273	0.166	0.099	0.166
Constant	−1.714	0.328	−0.917	0.300
ρ	0.924	0.027		
Log-likelihood	−255.782			

models is likely to be the DGP. Second, the models in the Monte Carlo experiment range from both being quite far from the DGP to one being far from the DGP and one quite close to the DGP. The two models in our application are some unknown distance away from the DGP. Third, the sample sizes in the Monte Carlo experiment range from 50 to 1000. The sample size in the application is 374. Fourth, and perhaps most importantly, the log-likelihood ratios in the experiment are leptokurtic, with kurtosis values ranging between 5.19 and 6.70. The log-likelihood ratios in the application are similarly leptokurtic, with a kurtosis value of 5.80. Fifth, the log-likelihood ratios in the experiment exhibit almost no skew. The same is true of the empirical log-likelihood ratios, which have a skewness of 0.028.[13] Thus, our empirical application is right in line with the experiment reported in Section 4.

Another question that might arise is the necessity of using a nonnested test in this situation. Given that the rival models are nonnested solely in terms of their design matrices, it might appear that neither the Vuong test nor the distribution-free test is necessary to discriminate between them. Conventional practice would suggest nesting the two models and employing a likelihood ratio test. As both Kmenta (1986, 596) and Greene (2003, 154) make clear with a simple example, the likelihood ratio test approach is not the appropriate method.

Consider two linear and additive models:

$$\text{Model 1} : \mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon},$$

$$\text{Model 2} : \mathbf{y} = \mathbf{Z}\gamma + \boldsymbol{\epsilon}.$$

Let $\tilde{\mathbf{X}}$ be the variables in $\mathbf{X}$, but not in $\mathbf{Z}$, $\tilde{\mathbf{Z}}$ be the variables in $\mathbf{Z}$, but not in $\mathbf{X}$, and finally, let $\mathbf{W}$ be the variables that are in both $\mathbf{X}$ and $\mathbf{Z}$. If we were to nest these models, the result would be $\mathbf{y} = \tilde{\mathbf{X}}\tilde{\beta} + \tilde{\mathbf{Z}}\tilde{\gamma} + \mathbf{W}\delta + \boldsymbol{\epsilon}$.

[13]We also bootstrapped the kurtosis and skewness of the empirical log-likelihoods. The bootstrap bias for kurtosis is −0.06, and the SE is 0.8. The bootstrap bias for skewness is −0.017, and the SE is 0.3.

Table 4 Results of the Vuong test for the Huth/Allee models

Vuong	*SE*	*Z statistics*	*Significance*	*95% confidence interval*	
−22.83	22.76	−1.00	0.316	−45.62	43.61

Notice that we cannot discriminate between model 1 and model 2 on the basis of a likelihood ratio test of either $\tilde{\beta} = \mathbf{0}$ or $\tilde{\gamma} = \mathbf{0}$ because these tests leave out δ, which is a mix of β and γ. Thus, the likelihood ratio test does not discriminate between model 1 and model 2, but rather, it discriminates between model 1 or model 2 and a hybrid model comprising the alternative and the null hypothesis. Just as in this simple example, the Huth and Allee models share a set of variables. Nesting them and using a likelihood ratio test, therefore, does not provide the answer for which we are looking.[14]

The results of the Vuong test for the Huth and Allee models are in Table 4.[15] The test returns a statistic of −22.83 and a confidence interval comfortably bracketing zero. We therefore fail to reject the null hypothesis of "no difference" at conventional significance levels and conclude that the models explain equally well. The direction of the coefficient, however, favors the political norms model.

The results of the distribution-free test for the Huth and Allee models are in Table 5. Where the Vuong test could not distinguish between the models, the distribution-free test readily distinguishes between them. We reject the null hypothesis of "no difference" in favor of the political norms model at conventional significance levels. Thus, the two tests are in the same direction, but only the distribution-free test actually rejects the null hypothesis. The fact that the distribution-free procedure can distinguish between these models given the complex nature of the estimator, the demands bivariate probit makes on the data, and the relatively small sample size make a strong statement about the utility of the procedure. In addition, the finding that a political norms explanation has greater explanatory power than a political structure explanation corroborates earlier work on the effect of domestic politics on foreign policy decision making (see Maoz and Russett 1993). Although this result does not overturn Huth and Allee's conclusions, which are based in part on empirical work on other stages besides escalation in international disputes, our result does suggest that more formal model testing could prove invaluable.

6 Conclusion

This paper considers a distribution-free alternative to the well-known Vuong test for nonnested model selection. Although both tests are consistent and unbiased, the distribution-free test is asymptotically more efficient for leptokurtic distributions. The Monte Carlo results demonstrate that in common research situations, such as choosing between competing models with nonnested design matrices, the distribution of individual log-likelihood ratios is actually leptokurtic. The results also confirm, as expected, that the distribution-free test has greater power than the Vuong test under these conditions. Thus,

[14] Even in the case of simple linear regression with no shared variables, the comprehensive approach is not recommended. No major econometrics text—Greene (2003), Kmenta (1986), Judge et al. (1985), Johnston and DiNardo (1997), Davidson and MacKinnon (1993)—presents it as a viable option. Other arguments against the comprehensive approach include low power (McAleer 1987), collinearity (Greene 2003), and the fact that the comprehensive model is atheoretical (Clarke 2001).

[15] *Stata* code for replicating the Huth and Allee tests is available in the supplementary material on the *Political Analysis* Web site. *R* code is available from the author.

Table 5 Results of the distribution-free test for the Huth/Allee models

One-sided tests
H_1: median of model $f - g > 0$
Binomial($n = 374$, $x \geq 149$, $p = 0.5$) = 1.0000
H_1: median of model $f - g < 0$
Binomial($n = 374$, $x \geq 225$, $p = 0.5$) = 0.0001
Two-sided tests
H_1: median of model $f - g \neq 0$
Min[1, 2 × Binomial($n = 374$, $x \geq 225$, $p = 0.5$)] = 0.0001

the new test is simple to perform, easy to interpret, and provides good power under difficult conditions.

As Vuong (1989, 326) 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 in terms of their functional forms as is being done by Clarke and Signorino (2006). Third, given that no particular correction for the number of parameters in the rival models can be justified, it would be useful 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. Fourth, comparisons to model selection criteria and the many Bayesian approaches to the problem would be enlightening. Finally, the extension of the distribution-free test to situations in which there are many competing models would be a greatly anticipated development.

References

Akaike, H. 1973. Information theory and an extension of the likelihood ratio principle. In *Second international symposium of information theory*, ed. B. N. Petrov and F. Csaki. Minnesota Studies in the Philosophy of Science, Budapest: Akademinai Kiado.

Albert, James H. 1996. Bayesian selection of log-linear models. *Canadian Journal of Statistics* 24:327–47.

Berger, James O., and Luis R. Pericchi. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91:109–122.

Brown, P. J., M. Vannucci, and T. Fearn. 1998. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* 60:627–41.

Cameron, A. Colin, and Pravin K. Trived. 2005. *Microeconometrics: Methods and applications.* Cambridge: Cambridge University Press.

Carlin, B. P., and S. Chibb. 1995. Bayesian model choice via Markov Chain Monte Carlo. *Journal of the Royal Statisitcal Society, Series B* 77:473–84.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch. 2001. The practical implementation of Bayesian model selection. In *Model selection*, ed. P. Lahiri. *Institute of mathematical statistics lecture notes.* Vol. 38, 67–116. Beachwood, OH: Institute of Mathematical Statistics.

Clarke, Kevin A. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45:724–74.

———. 2003. Nonparametric model discrimination in international relations. *Journal of Conflict Resolution* 47:72–93.

———. 2007. Data experiments: Model specifications as treatments. Unpublished manuscript.

Clarke, Kevin A., and Curt S. Signorino. 2006. Discriminating methods: Nonnested tests for strategic choice models. Unpublished manuscript.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and inference in econometrics.* Oxford: Oxford University Press.

Efron, Bradley. 1986. Why isn't everyone a Bayesian? *The American Statistician* 40:1–5.
Fernandez, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100:381–427.
George, Edward I., and Dean P. Foster. 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87:731–47.
Gibbons, Jean Dickinson, and Subhabrata Chakraborti. 1992. *Nonparametric statistical inference.* 3rd ed. New York: Marcel Dekker, Inc.
Greene, William H. 2003. *Econometric analysis.* 5th ed. Upper Saddle River, NJ: Prentice Hall.
Hodges, J. L., and E. L. Lehmann. 1956. The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics* 27:324–35.
Hollander, Myles, and Douglas A. Wolfe. 1999. *Nonparametric statistical methods.* 2nd ed. New York: John Wiley and Sons.
Huth, Paul K., and Todd Allee. 2002. *The democratic peace and territorial conflict in the twentieth century.* Cambridge studies in international relations, Cambridge: Cambridge University Press.
Johnston, Jack, and John DiNardo. 1997. *Econometric methods.* 4th ed. New York: McGraw Hill.
Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. 1985. *The theory and practice of econometrics.* 2nd ed. New York: John Wiley and Sons.
Kaiser, Henry F., and Kern Dickman. 1962. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika* 27:179–82.
Kmenta, Jan. 1986. *Elements of econometrics.* 2nd ed. New York: Macmillan Publishing Company.
Kullback, Solomon, and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.
Laud, Purushottam W., and Joseph G. Ibrahim. 1995. Predictive model selection. *Journal of the Royal Statisitcal Society, Series B* 57:247–62.
Lehmann, E. L. 1986. *Testing statistical hypotheses.* 2nd ed. New York: John Wiley and Sons.
Maoz, Zeev, and Bruce Russett. 1993. Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review* 87:624–38.
McAleer, Michael. 1987. Specification tests for separate models: A survey. In *specification analysis in the linear model*, ed. M. L. King and D. E. A. Giles. London: Routledge and Kegan Paul.
Noether, Gottfried E. 1955. On a theorem of Pitman. *Annals of Mathematical Statistics* 26:64–8.
———. 1967. *Elements of nonparametric statistics.* New York: John Wiley and Sons.
Pesaran, M. H. 1974. On the general problem of model selection. *Review of Economic Studies* 41:153–71.
———. 1987. Global and partial non-nested hypotheses and asymptotic local power. *Econometric Theory* 3:69–97.
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–4.
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* 57:307–33.
White, Halbert. 1982. Maximum likelihood estimator of misspecified models. *Econometrica* 50:1–25.