A Test for Global Maximum

Li GAN and Jiming JIANG

We give simple necessary and sufficient conditions for consistency and asymptotic optimality of a root to the likelihood equation. Based on the results, a large-sample test is proposed for detecting whether a given root is consistent and asymptotically efficient, a property often possessed by the global maximizer of the likelihood function. A number of examples, and the connection between the proposed test and the test of White for model misspecification, are discussed. Monte Carlo studies show that the test performs quite well when the sample size is large but may suffer the problem of overrejection with relatively small samples.

KEY WORDS: Asymptotic efficiency; Consistency; Large-sample test; Maximum likelihood; Multiple roots; Normal mixture.

1. INTRODUCTION

In many applications of the maximum likelihood method, people are bewildered by the fact that there may be multiple roots to the likelihood equation. Standard asymptotic theory asserts that under regularity conditions there exists a sequence of roots to the likelihood equation that is consistent but often gives no indication which root is consistent when the roots are not unique. Such results are often referred to as consistency of Cramér (Cramér 1946) type. In contrast, Wald (1949) proved that under some conditions, the global maximizer of the likelihood function is consistent. Because the global maximizer is typically a root to the likelihood equation, the so-called Wald consistency solves, from a theoretical viewpoint, the problem of identifying the consistent root. However, practically, the problem may still remain.

In practice, it is often much easier to find a root to the likelihood equation than the global maximizer of the likelihood function. In cases when there are multiple roots, if one could find all the roots, then, because the global maximizer is likely (but not surely) to be one of them, comparing the values of the likelihood at those roots would identify the global maximizer. But this strategy may still be impractical, because it may take considerable time to find a root, which makes it uneconomical or even impossible to find all of the roots. (We have not mentioned the fact that there are cases in which the likelihood equation may have an unbounded number of roots (see, e.g., Barnett 1966)). Quite often in practice, one, after some calculation, comes up with just one root. The question is whether this root is "a good one"; that is, the consistent root ensured by Cramér's theory.

It should be pointed out that in some rare cases even the global maximizer is not "good," and yet a local maximizer of the likelihood function that corresponds to a root to the likelihood equation may still be consistent (e.g., LeCam 1979; see also Lehmann 1983, ex. 6.3.1). Nevertheless, our main concern is to find a practical way of determining whether a given root is consistent. Note that under regularity conditions, consistency of a root to the likelihood equation implies asymptotic efficiency of it and vice versa (e.g., Lehmann 1983, sec. 6).

Previous studies of this problem are limited and concentrated on the method of evaluating a large number of likelihoods at different points of θ. For example, De Haan (1981) showed that a p-confidence interval of the global maximum (the largest of those likelihood values) can be constructed based on extreme-value asymptotic theory. Using this result, Veall (1991) conducted simulations of several econometric examples, each with at least two local optima. As Veall pointed out, this approach is essentially one method of grid searching. It requires a tremendous number of computations and often becomes impractical if the support of the parameter space is large and/or the parameter space is multidimensional.

In this article we give simple necessary and sufficient conditions for the consistency and asymptotic optimality of a root to the likelihood equation. The idea is based on a well-known fact that under regularity conditions, the logarithm of a likelihood function, say l, satisfies not only

$$E_{\theta_0}\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right) = 0, \tag{1}$$

but also

$$E_{\theta_0}\left(\left.\frac{\partial l}{\partial \theta}\right|_{\theta_0}\right)^2 + E_{\theta_0}\left(\left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\theta_0}\right) = 0, \tag{2}$$

where θ is an unknown parameter and θ_0 is the true θ. Equation (1) is consistent with the likelihood equation

$$\frac{\partial l}{\partial \theta} = 0. \tag{3}$$

The problem is that there may be multiple roots to (3): global maximizer, local maximizer, (local) minimizer, and so on. Our claim is very simple: A global maximizer would satisfy

$$\left(\frac{\partial l}{\partial \theta}\right)^2 + \frac{\partial^2 l}{\partial \theta^2} \approx 0, \tag{4}$$

Li Gan is Assistant Professor, Department of Economics, University of Texas, Austin, TX 78712 (E-mail: *gan@eco.utexas.edu*). Jiming Jiang is Assistant Professor, Department of Statistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: *jiang@eureka.cwru.edu*). The two authors contribute equally to this work. Computer programs related to the text are available at *http://www.eco.utexas.edu/~gan/global_max*. The authors thank Peter J. Bickel and Daniel L. McFadden for encouragement and suggestions, and Donald Andrews, Jianhua Huang, James Powell, Jiayang Sun, and Hal White for helpful discussion. The authors are also grateful to a referee and an associate editor for carefully reviewing the manuscript and for their insightful comments.

Journal of the American Statistical Association
September 1999, Vol. 94, No. 447, Theory and Methods

whereas an inconsistent root (which may correspond to a local maximizer or something else) would not. Note that (4) is consistent with (2). Therefore, this may be rephrased intuitively as that a "good" root is one that makes both (1) and (2) consistent with their counterparts, (3) and (4). To demonstrate this numerically, consider the following example.

Example 1 (Normal Mixture). A normal mixture density can be written as

$$f(x,\theta) = \frac{p}{\sigma_1}\,\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{1-p}{\sigma_2}\,\phi\left(\frac{x-\mu_2}{\sigma_2}\right), \quad (5)$$

where $\theta = \mu_1$ and $\phi(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$. Given σ_1, σ_2, and μ_2, the likelihood equation for θ typically has two roots if the two normal means are "well separated" (e.g., Titterington, Smith, and Makov 1985, sec. 5.5); one corresponds to a local maximum, and the other corresponds to a global maximum. Let the values be given as $\sigma_1 = 1, \mu_2 = 8, \sigma_2 = 4, p = .4$, and the true value of $\theta = -3$. We consider two ways of computing the standard errors: $\mathrm{SD}_1 = (E_\theta((\partial/\partial\theta)\log f(X,\theta))^2)^{-1/2}$ (the outer product form) and $\mathrm{SD}_2 = (-E_\theta(\partial^2/\partial\theta^2)\log f(X,\theta))^{-1/2}$ (the Hessian form). To see what might happen in practice, a random sample of size 5,000 was generated by a computer. The results are shown in Table 1. It is clear that for the global maximum, the two ways of computing the SD's give virtually the same result, indicating (4), whereas this is not true for the local maximum.

In Section 2 we make it clear mathematically what "≈ 0" means in (4), and prove that the claim is true under reasonable conditions. Based on the result, in Section 3 we propose a simple large-sample test for the consistency and asymptotic efficiency of the root and apply it to two examples.

It should be noted that the test used here is that given by White (1982). But while White used the test to detect model misspecification, we use it for another purpose: to detect whether a root corresponds to a global maximum. The basic idea of White's test is as follows: if (A) the model is correctly specified, and (B) $\hat\theta$ is the global maximizer of the log-likelihood, then one would expect $\phi_2(\hat\theta) \approx 0$, where $\phi_2(\theta) = (1/n)\sum_{i=1}^n[((\partial/\partial\theta)\log f(X_i,\theta))^2 + (\partial^2/\partial\theta^2)\log f(X_i,\theta)]$. In White's article (B) holds as an assumption and (A) is the null hypothesis. In this article the roles of (A) and (B) are reversed; that is, (A) holds as an assumption and (B) is the null hypothesis. White (1982) has given sufficient conditions for the asymptotic normality of $\sqrt{n}\phi_2(\hat\theta)$, which is the key to the test. In this article we establish necessary conditions for the asymptotic normality of White's test, which expand on the sufficient conditions given by White. In addition, one may also consider an estimate from local maxima as a form of misspecification, as implied in White's work (i.e., when one rejects, "one concludes ... as well as possible inconsistency of the QMLE for parameters of interest").

Table 1. Maxima in the Normal Mixture

	Global maximum	*Local maximum*
Starting values	−3.0	6.0
Estimated values	−2.9931	6.4778
Log-likelihood	−9901	−17443
SD_1	.0234	.0492
SD_2	.0237	.2807

A further question might be: What happens if one does not have knowledge about both (A) and (B)? This means that in testing for model misspecification, one does not know that $\hat\theta$, which is now a root to an estimating equation, corresponds to the global maximum of the object function, or when testing for global maximum, one does not know whether the model is correct. The answer is that in such a case, if the test rejects, then one concludes that there is either a problem of model misspecification or a problem of $\hat\theta$ not being a global maximum, but one cannot distinguish between these two. However, in many cases other techniques are available for model checking. For example, in the case of iid observations there are methods for formally (e.g., goodness-of-fit tests) or informally (e.g., Q-Q plots) checking the distributional model assumptions; in linear regression, model-diagnostic techniques are available for checking the correctness of the model. (For some recent development in model checking, see Jiang, Lahiri, and Wu 1998.) Once one is convinced that the model is correct, the only interpretation for the rejection is, of course, that $\hat\theta$ is not a global maximum. One thus searches for a new root, and then tests again for global maximum. In contrast, to the best of our knowledge no statistical method is available for checking whether a given root corresponds to a global maximum, even if the model has been determined correct. This is the main reason for writing this article.

2. CHARACTERISTICS FOR CONSISTENCY AND ASYMPTOTIC EFFICIENCY

Let $X_1, X_2, \ldots, X_n$ be independent with the same distribution that has a density function $f(x,\theta)$ with respect to some measure μ, where θ is a real-valued parameter. Suppose that the following regularity conditions hold:

a. The parameter space Θ is an open interval (not necessarily finite).
b. $f(x,\theta) \neq 0$ a.e. μ for all $\theta \in \Theta$.
c. $f(x,\theta)$ is three-times differentiable with respect to θ, and the integral $\int f(x,\theta)\,d\mu$ can be twice differentiated under the integral sign.

Let $\varphi_j(x,\theta) = (f(x,\theta))^{-1}(\partial^j/\partial\theta^j)f(x,\theta), j = 1, 2$. Let θ_0 be the true parameter. Hereafter we use $E(\cdot)$ for $E_{\theta_0}(\cdot)$ to simplify notation. We assume that the following:

d. $\max_{j=1,2} E|\varphi_j(X_1,\theta)| < \infty$ for all $\theta \in \Theta$, and $\max_{j=1,2} E\sup_{\theta\in\Theta\cap[-B,B]} |(\partial/\partial\theta)\varphi_j(X_1,\theta)| < \infty$ for all $B \in (0,\infty)$.

Let $d(\theta) = (d_1(\theta), d_2(\theta))$, where

$$d_j(\theta) = \int \varphi_j(x,\theta)f(x,\theta_0)\,d\mu, \qquad j = 1, 2. \quad (6)$$

Note that d may be regarded as a map from Θ to R^2, and it follows that $d^{-1}(\{(0,0)\}) \supseteq \{\theta_0\}$. Let $l(\theta) = \sum_{i=1}^n \log f(X_i,\theta)$ be the log-likelihood function based on observations $X_1, \ldots, X_n$, and let $\hat\theta$ be a root to (3). The following theorem gives necessary and sufficient conditions for the consistency of $\hat{\theta}$.

Theorem 1. Suppose that conditions a–d are satisfied. In addition, suppose that there is $M > 0$ such that

$$\{E(\inf_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta))\} \vee \{-E(\sup_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta))\} > 0 \quad (7)$$

($\inf_\emptyset\{\cdot\}$ and $\sup_\emptyset\{\cdot\}$ are understood as ∞ and $-\infty$), and that

$$d^{-1}(\{(0,0)\}) = \{\theta_0\}. \quad (8)$$

Then $\hat{\theta}\xrightarrow{P}\theta_0$ iff $P(\hat{\theta}\in\Theta)\to 1$ and

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_2(X_i,\hat{\theta}) \xrightarrow{P} 0. \quad (9)$$

Conditions a–d are similar to the regularity conditions of Lehmann (1983, p. 406): conditions a–c correspond to Lehmann's conditions (i)–(iv), and conditions d corresponds to Lehmann's condition (vi). Furthermore, (7) weakens the assumption of White (1982, assump. A2) that the parameter space is a compact subset of an Euclidean space. It is easy to see that if the parameter space is compact, then (7) is satisfied. For the most part, (7) says that the parameter space does not have to be compact, but the root to the equation

$$\left(\frac{\partial l}{\partial\theta}\right)^2 + \frac{\partial^2 l}{\partial\theta^2} = 0,$$

which corresponds to the information identity (2), lies asymptotically in a compact subspace. Note that

$$\varphi_2(X_i,\theta) = \left(\frac{\partial}{\partial\theta}\log f(X_i,\theta)\right)^2 + \frac{\partial^2}{\partial\theta^2}\log f(X_i,\theta),$$

and (7) means that either

$$E(\inf_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta)) > 0 \quad (10)$$

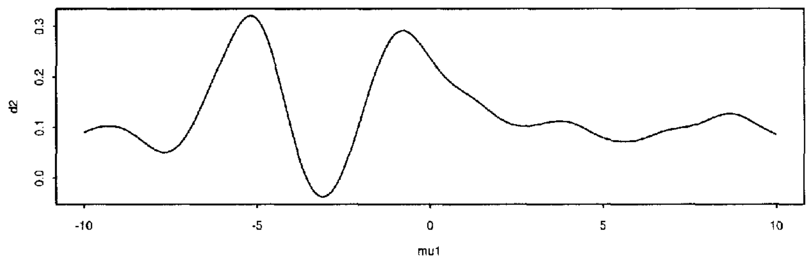
or

$$E(\sup_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta)) < 0. \quad (11)$$

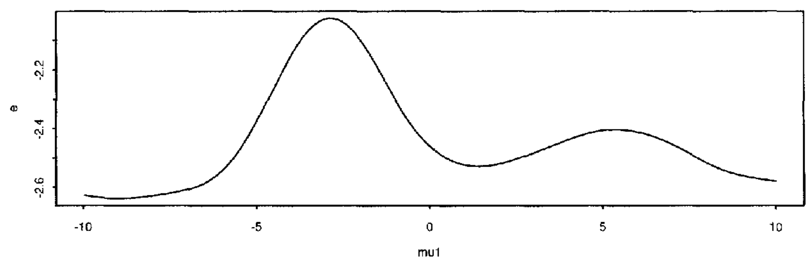
Intuitively, either (10) or (11) should hold, unless the density function $f(x,\theta)$ of the X_i's behaves strangely such that $\varphi_2(x,\theta)$, as a function of θ, changes sign infinitely many times outside any finite interval. Therefore, one would typically expect (7) to hold. To see a simple example, consider Example 3, which follows. It is shown that $\varphi_2(X_1,\theta) = (X_1 - e^\theta)^2 - e^\theta$. Thus it is easy to show that (10), and hence (7), hold (see the Appendix). Note that in this example the parameter space is not compact, $\Theta = (-\infty,\infty)$.

Perhaps the most notable condition in Theorem 1 is (8), which is an assumption on the distribution of $X_1,\ldots,X_n$; that is, on $f(x,\theta)$. We now use examples to further illustrate it.

Example 2 (Normal Mixture). Consider, again, the normal mixture distribution of Example 1. Although there is no analytic expression, the values of $d(\theta)$ can be computed by the Monte Carlo method. Figure 1 exhibits $d_2(\theta)$ and $e(\theta) = E\log f(X_1,\theta)$. It is seen that $e(\theta)$ has a global maximum at $\theta = \theta_0 = -3$ and a local maximum somewhere



(a)



(b)

Figure 1. Plots of (a) $d_2(\theta)$ and (b) $e(\theta)$ for the Normal Mixture With $\theta = \mu_1$ on the x-Axis.

between 5 and 10, which correspond to the two roots to $d_1(\theta) = 0$. However, only θ_0 coincides with the unique root to $d_2(\theta) = 0$. Therefore, (8) is satisfied.

Figure 1(a) might give the impression that $d_2(\theta)$ is always nonnegative, with 0 its minimum. This is not true, however, as is shown by the following example.

Example 3 (Poisson). Suppose that $X_1 \sim \text{Poisson}(e^\theta)$. Then $f(x,\theta) = (e^\theta)^x \exp(-e^\theta)/x! = \exp(\theta x - e^\theta - \log x!), x = 0, 1, 2, \ldots$ Thus $(\partial/\partial\theta) f(x,\theta) = f(x,\theta)(x - e^\theta)$ and $(\partial^2/\partial\theta^2) f(x,\theta) = f(x,\theta)[(x - e^\theta)^2 - e^\theta]$. Therefore, $d_2(\theta) = E\varphi_2(X_1, \theta) = \text{var}(X_1) + (e^{\theta_0} - e^\theta)^2 - e^\theta = e^{\theta_0} - e^\theta + (e^{\theta_0} - e^\theta)^2$, which can be negative. For example, if $\theta_0 = 0, \theta = .2$, then $d_2(\theta) \approx -.1724$. In fact, in this example $d_2(\theta) = 0$ has two roots: θ_0 and $\log(1 + e^{\theta_0})$, whereas $d_1(\theta) = 0$ has a single root θ_0. Note that the situation here is contrary to that in Example 2, in which $d_1(\theta) = 0$ has two roots and $d_2(\theta) = 0$ has one. But the bottom line is that in both cases, (8) is satisfied.

The next example is selected from an econometrics book.

Example 4 (Amemiya 1994). Suppose that $X_1 \sim N(\theta, \theta^2)$. Then direct calculation shows that

$$d_1(\theta) = -\frac{1}{\theta} - \frac{\theta_0}{\theta^2} + \frac{2\theta_0^2}{\theta^3}$$

and

$$d_2(\theta) = \frac{2}{\theta^2} + \frac{4\theta_0}{\theta^3} - \frac{8\theta_0^2}{\theta^4} - \frac{8\theta_0^3}{\theta^5} + \frac{10\theta_0^4}{\theta^6}.$$

It is easy to show that $d_1(\theta) = 0$ has two roots: θ_0 and $-2\theta_0$. However, we have $d_2(-2\theta_0) = -3/32\theta_0^2 < 0$. Therefore, (8) is, again, satisfied.

After seeing three examples, some may conjecture that perhaps (8) is always satisfied. This turns out, again, to be wrong.

Example 5 (A Counterexample). Let $\theta_0 < \theta_1$ be fixed. Let

$$\varphi(\theta) = \frac{1}{4}\left[1 + \frac{\theta - \theta_0}{3(\theta_1 - \theta_0)}\right]$$

and

$$\psi(\theta) = \frac{1}{6}\left[2 + \frac{(\theta_1 - \theta)^3}{(\theta_1 - \theta_0)^3}\right].$$

It is easy to show that $\frac{7}{12} \le \varphi(\theta) + \psi(\theta) \le \frac{3}{4}$. Therefore, there is $\varepsilon > 0$ such that $\varphi(\theta), \psi(\theta) > 0$, and $\varphi(\theta) + \psi(\theta) < 1, \theta_0 - \varepsilon \le \theta \le \theta_1 + \varepsilon$. Let $A = \theta_0 - \varepsilon, B = \theta_1 + \varepsilon$.

Let $a < b < c$ be real numbers. Let the parameter space for θ be (A, B), and let θ_0 be the true parameter. Define $f(a,\theta) = \varphi(\theta), f(b,\theta) = \psi(\theta)$, and $f(c,\theta) = 1 - \varphi(\theta) - \psi(\theta)$. Then for any $\theta \in (A, B), f(\cdot,\theta)$ is a probability density on $S = \{a, b, c\}$ with respect to counting measure μ [i.e., $\mu(\{x\}) = 1$ if $x \in S$ and 0 if $x \notin S$]; and for each $x \in S, f(x,\cdot)$ is twice continuously differentiable with respect to $\theta \in (A, B)$. However, it is easy to show that $d_j(\theta_1) = 0, j = 1, 2$. Thus (8) is not satisfied. Furthermore, it can be shown that

$$E_{\theta_0}\left(\frac{\partial^2}{\partial\theta^2}\log f(X,\theta_1)\right) = -\frac{9}{2}\left(\frac{\partial}{\partial\theta} f(a,\theta_1)\right)^2 < 0.$$

Therefore, θ_1 corresponds to a local maximum of the (expected) log-likelihood.

This example also shows that for the conclusion of Theorem 1 to hold, (8) cannot be dropped. To see this, let, for simplicity, $\theta_j = j, j = 0, 1$. Let $X_1, \ldots, X_n$ be an iid sample from $f(\cdot,\theta_0)$. Then it can be shown that with probability tending to 1 there is a root $\hat\theta_1$ to the likelihood equation such that $\hat\theta_1 \xrightarrow{P} \theta_1 \ne \theta_0$, although all the conditions of Theorem 1 except (8) are satisfied and $(1/n)\sum_{i=1}^n \varphi_2(X_i, \hat\theta_1) \xrightarrow{P} 0$ (see the Appendix).

Note. Although the counterexample shows that Theorem 1 cannot hold without (8), it does not necessarily imply limitation on the applicability of the test that we propose in the next section. This is because Example 5 is rather artificial. In most practical situations, (8) is expected to hold. Also note that (8) is needed only for the "if" part of Theorems 1 and Theorem 2 (see the proof in the Appendix). Furthermore, the idea of our test for global maximum comes from the observation that the following hold when $\theta = \theta_0$:

$$d_j(\theta) \equiv E_{\theta_0}\left[\frac{f^{(j)}(X_1,\theta)}{f(X_1,\theta)}\right] = 0, \qquad (12)$$

$j = 1, 2$, where $f^{(j)}(x,\theta) = (\partial^j/\partial\theta^j) f(x,\theta)$. However, as Example 5 shows, there may be $\theta \ne \theta_0$ that satisfies (12). This is why condition (8) is needed. Note that (8) means that

$$d_j(\theta) = 0, j = 1, 2 \quad \text{iff } \theta = \theta_0. \qquad (13)$$

More generally, note that in fact $\theta = \theta_0$ satisfies a series of equations; that is, (12) with $j = 1, 2, 3, \ldots$ The implication is that, for example, a distribution not satisfying (13) may actually satisfy

$$d_j(\theta) = 0, \qquad j = 1, 2, 3 \quad \text{iff } \theta = \theta_0. \qquad (14)$$

For example, it is easy to show that for the distribution in Example 5, which breaks (13), $d_3(\theta_1) = -\frac{3}{4}(\theta_1 - \theta_0)^3 \ne 0$, and thus (14) is satisfied. Thus one may consider a generalized test based on $\varphi_3(X_i, \hat\theta)$ (defined likewise) instead of $\varphi_2(X_i, \hat\theta)$, and expect the conclusions of Theorems 1 and Theorem 2 to hold for a broader class of distributions [i.e., replacing (13) by (14)]. In related work, Andrews (1997) proposed a stopping rule for global minimum in the context of generalized method of moments assuming only one root in the moment conditions. Theorem 1 and (14) in this paper may be considered as extra moment conditions to ensure that such an assumption holds and his stopping rule may then be used. (We thank Don Andrews for his insights on this point.)

Let $I(\theta) = E_\theta(\varphi_1(X_1,\theta))^2$ be the Fisher information, which equals $-E_\theta(\partial^2/\partial\theta^2)\log f(X_1,\theta)$ under condition c. We now assume, in addition to conditions a–d, the following:

e. $0 < I(\theta_0) < \infty$.

f. $f(x, \theta)$ is four-times differentiable with respect to θ, and there is $\delta > 0$ such that $\max_{j=1,2} E \sup_{\theta \in \Theta \cap [\theta_0 - \delta, \theta_0 + \delta]} |(\partial^2/\partial\theta^2)\varphi_j(X_1, \theta)| < \infty$.

Let $\phi_j(\theta) = (1/n)\sum_{i=1}^{n} \varphi_j(X_i, \theta)$ and $\psi(x, \theta) = \varphi_2(x, \theta) - (E_\theta \phi_2'(\theta)/E_\theta \phi_1'(\theta))\varphi_1(x, \theta)$ whenever $E_\theta \phi_1'(\theta) = -I(\theta) \neq 0$. The following theorem gives necessary and sufficient conditions for the asymptotic efficiency of $\hat{\theta}$.

Theorem 2. Suppose that conditions a–f are satisfied and that (7) (for some $M > 0$) and (8) hold. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathrm{N}(0, I(\theta_0)^{-1}) \tag{15}$$

iff $P(\hat{\theta} \in \Theta) \to 1$ and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_2(X_i, \hat{\theta}) \xrightarrow{\mathcal{L}} \mathrm{N}(0,\ \mathrm{var}(\psi(X_1, \theta_0))). \tag{16}$$

In addition to the assumptions explained earlier, conditions e and f are similar to the regularity conditions (v) and (vi) of Lehmann (1983, p. 406; see also thm. 2.3 therein).

An equivalence of the "only if" part of Theorem 2 has been obtained by White (1982, thm. 4.1), which states that under suitable conditions, the normalized test statistic, which corresponds to the square of our test statistic in the univariate case, has an asymptotic chi-squared distribution. Despite the fact that the two theorems deal with different problems, an obvious contribution of Theorem 2 is that it gives a necessary and sufficient condition for the asymptotic normality, whereas theorem 4.1 of White (1982) provides only sufficient conditions.

Also, the conditions of Theorem 2 are somehow more general than those of White's theorem 4.1. The latter requires that the parameter space be a compact subset of an Euclidean space, which rules out some important cases where the parameter space is unbounded (e.g., the mean of a normal distribution). Theorem 2 weakens such a condition to (7). As discussed earlier, this allows our theorem to have broader applicability.

The proofs of Theorems 1 and 2 are given in the Appendix.

3. A LARGE-SAMPLE TEST AND MONTE CARLO STUDIES

Let $s(\theta_0) = \sqrt{\mathrm{var}(\psi(X_1, \theta_0))}$. Assuming that $s(\cdot)$ is continuous, (16) implies that

$$\sqrt{n}\phi_2(\hat{\theta})/s(\hat{\theta}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, 1). \tag{17}$$

Therefore, the statistic $\sqrt{n}\phi_2(\hat{\theta})/s(\hat{\theta})$ may be used for a large-sample test for consistency and asymptotic efficiency of the root $\hat{\theta}$. Such a statistic has been suggested by White (1982) for (large-sample) testing for model misspecification. A further expression for $s(\hat{\theta})$ is given as follows. Because $E\varphi_j(X_1, \theta_0) = 0, j = 1, 2$, we have

$$\begin{aligned}
(s(\theta_0))^2 &= E(\psi(X_1, \theta_0))^2 \\
&= E(\varphi_2(X_1, \theta_0))^2 + 2\,\frac{E\phi_2'(\theta_0)}{I(\theta_0)} \\
&\quad \times E\varphi_1(X_1, \theta_0)\varphi_2(X_1, \theta_0) + \left(\frac{E\phi_2'(\theta_0)}{I(\theta_0)}\right)^2 \\
&\quad \times E(\varphi_1(X_1, \theta_0))^2 \\
&= E(\varphi_2(X_1, \theta_0))^2 + I(\theta_0)^{-1} \\
&\quad \times \left[2E\varphi_1(X_1, \theta_0)\varphi_2(X_1, \theta_0) E\,\frac{\partial}{\partial\theta}\,\varphi_2(X_1, \theta)|_{\theta_0}\right. \\
&\quad \left. + \left(E\,\frac{\partial}{\partial\theta}\,\varphi_2(X_1, \theta)|_{\theta_0}\right)^2\right].
\end{aligned} \tag{18}$$

(Note that all of the expectations are taken at θ_0.) By taking the square root and replacing θ_0 by $\hat{\theta}$, we get an expression for $s(\hat{\theta})$.

In practice, it is sometimes more convenient to use an alternative statistic that also has an asymptotic N(0, 1) distribution: $\phi_2(\hat{\theta})/\mathrm{SD}(\phi_2(\hat{\theta}))$. The advantage of this approach is that the denominator can be approximated by the Monte Carlo method (i.e., the *bootstrap*).

To illustrate the test, we carry out Monte Carlo studies using two previously discussed examples, Examples 1 and 4.

Example 1 Revisited. This example deals with the normal mixture that has been frequently used (e.g., Hamilton 1989). It is also one of the classical situations in which the likelihood equation may have multiple roots. For these reasons, we conduct a simulation study on the performance of the large-sample test proposed in the case of Example 1.

We consider sample sizes n = 1,000, 500, and 250. In many economic problems the sample sizes are much larger; thus even n = 1,000 is not impractical. For each sample size we simulated 500 datasets from a normal mixture distribution with parameters $\mu_1 = -3, \mu_2 = 8, \sigma_1 = 1, \sigma_2 = 4$, and $p = .4$. For each dataset we applied the test to both the global and the local maximizers (of the log-likelihood) found. Table 2 reports the observed significance level and power at the alternative (i.e., the local maximizer) at significance levels $\alpha = .05$ and $.10$.

Table 2. Level and Power for Normal Mixture

	$\mu_2 = 8$						$\mu_2 = 2$	
	n = 1,000		n = 500		n = 250		n = 500	
	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)
α = .05	.05	1.00	.05	.94	.05	.43	.05	.97
α = .10	.11	1.00	.13	.99	.11	.65	.13	.99

Table 3. Level and Power for Example 4 (Amemiya 1994)

	n = 1,000		n = 500		n = 250	
	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)
α = .05	.064	1.00	.066	.980	.104	.866
α = .10	.098	1.00	.122	.998	.162	.908

It is seen that for all sample sizes, the observed significance level is identical or very close to α. For $n = 1{,}000$ or 500, the power is either 1.0 or very close. On the other hand, the power drops dramatically for $n = 250$. Some may wonder whether the very high power (except for $n = 250$) is due to the fact that $\mu_2 = 8$ is far away from $\mu_1 = -3$. To find out the answer, we change μ_2 from 8 to 2 (and leave others unchanged) and repeat the simulation for $n = 500$; the result is also summarized in Table 2. It might be worth pointing out that when $\mu_2 = 2$, one cannot always find the local maximizer. In fact, in 314 of 500 cases we are able to find the local maximizer, and hence we report the observed power based on those 314.

Example 4 Revisited. Consider Example 4, which comes from Amemiya (1994), $N(\theta, \theta^2)$. Table 3 gives the level and power with 500 replications. We use $\theta_0 = 1$ to generate the data. In this case, when the number of observations is relatively large, one has the correct level and very high power. The test suffers the problem of overrejection at relatively smaller samples. The small-sample property of this test certainly deserves further investigation.

4. CONCLUDING REMARKS

Although in this article we have assumed that θ is a real-valued parameter, there is no essential difficulty to generalize the results to the case where θ is multidimensional. The property on which we rely in this test is the information matrix equality, which holds asymptotically at the global maximum. The size of the test will always be asymptotically correct given a correctly specified model. The consistency of the test depends on condition (8). A rather artificial counterexample is constructed to show that condition (8) cannot be dropped universally. However, we also illustrate that a modification of (8) based on higher derivatives is potentially useful.

APPENDIX: PROOFS

Proof That (10) Holds for Example 3

Let $a = \inf_{\theta\in[-M,M]^c} \varphi_2(X_1,\theta)$. Suppose that M is suitably large. Then, if $1 \le X_1 \le (\frac{1}{2})e^M$, we have $\varphi_2(X_1,\theta) \ge (1-e^\theta)^2 - e^\theta > 1 - 3e^{-M}, \theta < -M$, and $\varphi_2(X_1,\theta) > (e^\theta/2)^2 - e^\theta \ge (\frac{1}{4})e^{2M} - e^M, \theta > M$. Therefore, $a \ge (1-3e^{-M}) \wedge ((\frac{1}{4})e^{2M} - e^M)$. If $X_1 = 0$, then $\varphi_2(X_1,\theta) = e^{2\theta} - e^\theta > e^{2M} - e^M > 0$, $\theta > M$, and $\varphi_2(X_1,\theta) > -e^\theta > -e^{-M}, \theta < -M$. Thus $a \ge -e^{-M}$. Finally, suppose that $X_1 > (\frac{1}{2})e^M$ and $\theta > M$. If $\varphi_2(X_1,\theta) \le 0$, then $e^\theta \le X_1 + |X_1 - e^\theta| \le X_1 + e^{\theta/2} \le X_1 + (\frac{1}{2})e^\theta$. Thus $e^\theta \le 2X_1$, and hence $|\varphi_2(X_1,\theta)| \le 4X_1^2 + 2X_1$. Therefore, $\varphi_2(X_1,\theta) \ge -(4X_1^2 + 2X_1)$. Now, suppose that $\theta < -M$. Then $|\varphi_2(X_1,\theta)| < (X_1 \vee 1)^2 + 1$. Therefore, $\varphi_2(X_1,\theta) > -((X_1 \vee 1)^2 + 1)$. Thus we have $a > -[(4X_1^2 + 2X_1) \vee ((X_1 \vee 1)^2 + 1)]$. It follows that

$$
\begin{aligned}
Ea &= Ea1_{(X_1=0)} + Ea1_{(1\le X_1 \le (1/2)e^M)} + Ea1_{(X_1>(1/2)e^M)} \\
&\ge -e^{-M}P(X_1=0) + (1-3e^{-M}) \wedge \left(\frac{1}{4}\,e^{2M} - e^M\right) \\
&\quad \times P\left(1 \le X_1 \le \frac{1}{2}\,e^M\right) \\
&\quad - [(4X_1^2 + 2X_1) \vee ((X_1 \vee 1)^2 + 1)]P\left(X_1 > \frac{1}{2}\,e^M\right) \to \\
&\quad P(X_1 \ge 1) > 0,
\end{aligned}
$$

as $M \to \infty$. Therefore, for suitably large M, (10) is satisfied.

Proof of Theorem 1

First, we note that $d_j(\theta) = E\varphi_j(X_1,\theta)$. Without loss of generality, assume that $\Theta \cap [-M,M]^c \ne \emptyset$ and thus contains a point θ_*. Then (7) implies that either $0 < \delta_1 = E\inf_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_1,\theta) \le E\varphi_2(X_1,\theta_*) < \infty$ or $0 > \delta_2 = E\sup_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_1,\theta) \ge E\varphi_2(X_1,\theta_*) > -\infty$.

Suppose that $\hat\theta \xrightarrow{P} \theta_0$. Then, by condition a, we have $P(\hat\theta \in \Theta) \to 1$. By condition c and the strong law of large numbers (SLLN), we have $\phi_2(\theta_0) \xrightarrow{\text{a.s.}} 0$. Also, by condition d and Taylor expansion, we have, when $\hat\theta \in \Theta$ and $|\hat\theta - \theta_0| \le \delta$, that

$$
\begin{aligned}
&|\phi_2(\hat\theta) - \phi_2(\theta_0)| \\
&\quad \le |\hat\theta - \theta_0| \left(\frac{1}{n}\sum_{i=1}^n \sup_{\theta\in\Theta\cap[\theta_0-\delta,\theta_0+\delta]} |(\partial/\partial\theta)\varphi_2(X_i,\theta)|\right) \\
&\quad = |\hat\theta - \theta_0| O_P(1).
\end{aligned}
$$

Therefore, (9) holds.

Now suppose that $P(\hat\theta \in \Theta) \to 1$ and (9) holds. We first show that

$$
P(|\hat\theta| > M) \to 0. \tag{A.1}
$$

In fact, suppose that $0 < \delta_1 < \infty$; then $\hat\theta \in \Theta$ and $|\hat\theta| > M$ imply

$$
\phi_2(\hat\theta) \ge \frac{1}{n}\sum_{i=1}^n \xi_i \xrightarrow{\text{a.s.}} \delta_1,
$$

where $\xi_i = \inf_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_i,\theta)$. Thus

$$
\begin{aligned}
P(|\hat\theta| > M) &\le P\left(\frac{1}{n}\sum_{i=1}^n \xi_i \le \frac{\delta_1}{2}\right) \\
&\quad + P\left(|\hat\theta| > M, \hat\theta \in \Theta, \frac{1}{n}\sum_{i=1}^n \xi_i > \frac{\delta_1}{2}\right) \\
&\quad + P(\hat\theta \notin \Theta) \le P\left(\frac{1}{n}\sum_{i=1}^n \xi_i \le \frac{\delta_1}{2}\right) \\
&\quad + P\left(\phi_2(\hat\theta) > \frac{\delta_1}{2}\right) + P(\hat\theta \notin \Theta) \to 0.
\end{aligned}
$$

Similarly, one can show (A.1) provided that $0 > \delta_2 > -\infty$.

Suppose that for some $\delta_0 > 0$, it is not true that $P(|\hat\theta - \theta_0| \geq \delta_0) \to 0$. Without loss of generality, suppose that

$$P(|\hat\theta - \theta_0| \geq \delta_0) \geq \varepsilon_0 \qquad \text{(A.2)}$$

for some $\varepsilon_0 > 0$ and all n. Let $\phi(\theta) = (\phi_1(\theta), \phi_2(\theta))$. Then condition d implies that $E\phi(\cdot)$ is continuous. Let

$$B_j = E\left(\sup_{\theta\in\Theta, |\theta|\leq M+2|\theta_0|}\left|\frac{\partial}{\partial\theta}\varphi_j(X_1,\theta)\right|\right), \qquad j = 1, 2.$$

Because $\theta_0 \notin \Theta_1 = \{\theta \in \Theta\colon \delta_0 \leq |\theta - \theta_0| \leq M + |\theta_0|\}$, we have, by (8), that $\rho = \inf_{\theta\in\Theta_1}|E\phi(\theta)| > 0$. Let $\eta = \rho/4(B_1 \vee B_2)$. Then there are an integer m and points $\theta_1, \theta_2, \ldots, \theta_m \in \Theta_1$ such that for any $\theta \in \Theta_1$ there is $1 \leq l \leq m$ such that $|\theta_l - \theta| < \eta$. Suppose that $\hat\theta \in \Theta, |\hat\theta| \leq M$, and $|\hat\theta - \theta_0| \geq \delta_0$. Then $\hat\theta \in \Theta_1$, and hence there is $\tilde\theta \in \{\theta_l, 1 \leq l \leq m\}$ such that $|\tilde\theta - \hat\theta| < \eta$. Thus by Taylor expansion and SLLN,

$$\begin{aligned}
|\phi_2(\hat\theta)| &= |\phi(\hat\theta)| \\
&\geq |\phi(\tilde\theta)| - |\phi(\tilde\theta) - \phi(\hat\theta)| \\
&\geq \min_{1\leq l\leq m}|\phi(\theta_l)| - |\phi_1(\tilde\theta) - \phi_1(\hat\theta)| - |\phi_2(\tilde\theta) - \phi_2(\hat\theta)| \\
&\geq \frac{\rho}{2} - \max_{1\leq l\leq m}|\phi(\theta_l) - E\phi(\theta_l)| \\
&\quad - \eta\sum_{j=1}^{2}\left(\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta,|\theta|\leq M+2|\theta_0|}\left|\frac{\partial}{\partial\theta}\varphi_j(X_i,\theta)\right| - B_j\right) \\
&= \frac{\rho}{2} - o_P(1).
\end{aligned}$$

Therefore, for the same $o_P(1)$,

$$\begin{aligned}
&P(|\phi_2(\hat\theta)| \geq \rho/4) \\
&\quad\geq P(|\phi_2(\hat\theta)| \geq \rho/2 - o_P(1), |o_P(1)| \leq \rho/4) \\
&\quad\geq P(|\phi_2(\hat\theta)| \geq \rho/2 - o_P(1)) - P(|o_P(1)| > \rho/4) \\
&\quad\geq P(\hat\theta \in \Theta, |\hat\theta| \leq M, |\hat\theta - \theta_0| \geq \delta_0) - P(|o_P(1)| > \rho/4) \\
&\quad\geq P(|\hat\theta - \theta_0| \geq \delta_0) - P(\hat\theta \notin \Theta) - P(|\hat\theta| > M) \\
&\qquad - P(|o_P(1)| > \rho/4) \\
&\quad\geq \varepsilon_0 - o(1) \to \varepsilon_0
\end{aligned}$$

as $n \to \infty$, which contradicts (9). Therefore, we must have $\hat\theta \xrightarrow{P} \theta_0$.

Proof of Theorem 2

Suppose that (15) holds. Then, by condition a, $P(\hat\theta \in \Theta) \to 1$. By Taylor series expansion, we have

$$0 = \phi_1(\hat\theta) = \phi_1(\theta_0) + \phi_1'(\theta_0)(\hat\theta - \theta_0) + \frac{1}{2}\,\phi_1''(\theta_1)(\hat\theta - \theta_0)^2,$$

where θ_1 is between θ_0 and $\hat\theta$. This implies

$$\sqrt{n}(\hat\theta - \theta_0) = -(\phi_1'(\theta_0) + \frac{1}{2}\,\phi_1''(\theta_1)(\hat\theta - \theta_0))^{-1}\sqrt{n}\phi_1(\theta_0). \quad \text{(A.3)}$$

On the other hand, we have

$$\phi_2(\hat\theta) = \phi_2(\theta_0) + \phi_2'(\theta_0)(\hat\theta - \theta_0) + \frac{1}{2}\,\phi_2''(\theta_2)(\hat\theta - \theta_0)^2, \quad \text{(A.4)}$$

where θ_2 is between θ_0 and $\hat\theta$. Combining (A.3) and (A.4), we have

$$\begin{aligned}
\sqrt{n}\phi_2(\hat\theta) &= \left[\sqrt{n}\phi_2(\theta_0) - \frac{E\phi_2'(\theta_0)}{E\phi_1'(\theta_0)}\sqrt{n}\phi_1(\theta_0)\right] \\
&\quad - \left(\frac{\phi_2'(\theta_0)}{\phi_1'(\theta_0) + (\frac{1}{2})\phi_1''(\theta_1)(\hat\theta - \theta_0)} - \frac{E\phi_2'(\theta_0)}{E\phi_1'(\theta_0)}\right) \\
&\quad \times \sqrt{n}\phi_1(\theta_0) + \frac{1}{2}\phi_2''(\theta_2)\sqrt{n}(\hat\theta - \theta_0)^2. \qquad \text{(A.5)}
\end{aligned}$$

By SLLN, $\phi_j'(\theta_0) \xrightarrow{\text{a.s.}} E\phi_j'(\theta_0), j = 1, 2$. Let $S = \{\hat\theta \in \Theta, |\hat\theta - \theta_0| \leq \delta\}$. Then condition a and (15) imply that $P(S) \to 1$. Because on S we have

$$|\phi_1''(\theta_1)| \leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta,|\theta-\theta_0|\leq\delta}\left|\frac{\partial^2}{\partial\theta^2}\varphi_1(X_i,\theta)\right|, \qquad \text{(A.6)}$$

and, by condition f, the right side of (A.6) is bounded in L^1, we have $\phi_1''(\theta_1)(\hat\theta - \theta_0) = o_P(1)$. These, combined with the fact that $E(\sqrt{n}\phi_1(\theta_0))^2 = I(\theta_0)$ and condition e, imply that the second term on the right side of (A.5) is $o_P(1)$. Similarly, the third term on the right side of (A.5) $= (\frac{1}{2})\phi_2''(\theta_2)\cdot\sqrt{n}(\hat\theta - \theta_0)\cdot(\hat\theta - \theta_0) = o_P(1)$. Finally, by the central limit thoerem (CLT), the first term on the right side of (A.5) $= (1/\sqrt{n})\sum_{i=1}^{n}\psi(X_i,\theta_0) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \mathrm{var}(\psi(X_1,\theta_0)))$. Therefore, (16) holds.

Now supose that $P(\hat\theta \in \Theta) \to 1$ and (16) holds. Then $\phi_2(\hat\theta) \xrightarrow{P} 0$, and hence by Theorem 1, $\hat\theta \xrightarrow{P} \theta_0$. On the other hand, we have by CLT that $\sqrt{n}\phi_1(\theta_0) \xrightarrow{\mathcal{L}} \mathrm{N}(0, I(\theta_0))$. Thus (15) follows by (A.3).

A Proof Regarding Example 5

Let $I_x = \{1 \leq i \leq n\colon X_i = x\}$, and $n_x = |I_x|$($|\cdot|$ denotes cardinality), $x \in S$. Then it is easy to show that the likelihood equation is equivalent to

$$\begin{aligned}
\hat s(\theta) = \left(\frac{n_a}{n}\right)\frac{\varphi'(\theta)}{\varphi(\theta)} + \left(\frac{n_b}{n}\right)\frac{\psi'(\theta)}{\psi(\theta)}& \\
- \left(1 - \frac{n_a}{n} - \frac{n_b}{n}\right)\frac{\varphi'(\theta) + \psi'(\theta)}{1 - \varphi(\theta) - \psi(\theta)} &= 0. \quad \text{(A.7)}
\end{aligned}$$

On the other hand, we have $\hat s(\theta) \xrightarrow{\text{a.s}} s(\theta)$ for fixed θ, where

$$s(\theta) = \frac{1}{4}\frac{\varphi'(\theta)}{\varphi(\theta)} + \frac{1}{2}\frac{\psi'(\theta)}{\psi(\theta)} - \frac{1}{4}\frac{\varphi'(\theta) + \psi'(\theta)}{1 - \varphi(\theta) - \psi(\theta)}. \qquad \text{(A.8)}$$

Let $\theta = 1 + x$. Then it is seen that for small $|x|$,

$$s(1 + x) = x\left[\frac{1 - x^2}{2(4 + x)(4 - x + 2x^3)} + o(1)\right]. \qquad \text{(A.9)}$$

Therefore, there is $0 < \delta_0 < 1$ such that for any $0 < \delta \leq \delta_0$, we have $s(1 - \delta) > 0$ and $s(1 + \delta) < 0$. Now let $0 < \delta \leq \delta_0$ be fixed. By (A.7), $\hat s(1 - \delta) \xrightarrow{\text{a.s}} s(1 - \delta)$, $\hat s(1 + \delta) \xrightarrow{\text{a.s.}} s(1 + \delta)$; thus

$$\begin{aligned}
&P_{\theta_0}(\exists \text{ root to } \hat s(\theta) = 0 \text{ in } (1 - \delta, 1 + \delta)) \\
&\quad\geq P_{\theta_0}\left(\hat s(1 - \delta) \geq \frac{s(1 - \delta)}{2}, \hat s(1 + \delta) \leq \frac{s(1 + \delta)}{2}\right) \to 1.
\end{aligned} \qquad \text{(A.10)}$$

Let $\hat\theta_1$ be a root to $\hat s(\theta) = 0$ that is within $(1 - \delta, 1 + \delta)$. (If there are more than one, let $\hat\theta_1$ be the one closest to 1.) Then, by (A.10) and the arbitrariness of δ, we have $\hat\theta_1 \xrightarrow{P} \theta_1 = 1 \neq \theta_0$.

It is obvious that conditions a–d and (7) are satisfied. However, it is easy to show that as $\theta \to 1, (\partial/\partial\theta)\varphi_2(x,\theta)$ is bounded for any $x \in S$. Therefore, by Taylor expansion,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\varphi_2(X_i,\hat{\theta}_1)\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\varphi_2(X_i,\theta_1) + \frac{1}{n}\sum_{i=1}^{n}[\varphi_2(X_i,\hat{\theta}_1) - \varphi_2(X_i,\theta_1)]\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\varphi_2(X_i,\theta_1) + (\hat{\theta}_1 - \theta_1)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\varphi_2(X_i,\tilde{\theta}_i)\right)\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\varphi_2(X_i,\theta_1) + (\hat{\theta}_1 - \theta_1)O_p(1),
\end{aligned}$$

where $\tilde{\theta}_i$ is between θ_1 and $\hat{\theta}_1$. Thus $(1/n)\sum_{i=1}^{n}\varphi_2(X_i,\hat{\theta}_1) \xrightarrow{P} E_{\theta_0}\varphi_2(X_1,\theta_1) = d_2(\theta_1) = 0$.

[Received July 1997. Revised March 1999.]

REFERENCES

Amemiya, T. (1994), *Introduction to Statistics and Econometrics*, Cambridge, MA: Harvard University Press.

Andrews, D. (1997), "A Stopping Rule for the Computation of Generalized Method-of-Moments Estimation," *Econmometrica*, 65, 913–931.

Barnett, V. D. (1966), "Evaluation of the Maximum Likelihood Estimator Where the Likelihood Equation has Multiple Roots," *Biometrika*, 53, 151–166.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.

De Haan, L. (1981), "Estimation of the Minimum of a Function Using Order Statistics," *Journal of the American Statistical Association*, 76, 467–469.

Hamilton, J. B. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

Jiang, J., Lahiri, P., and Wu, C. (1988), "On Pearson χ^2 Testing With Unobservable Cell Frequencies and Mixed Model Diagnostics," technical report, Case Western Reserve University, Department of Statistics.

LeCam, L. (1979), "Maximum Likelihood: An Introduction, Lecture Notes in Statistics 18, College Park, MD: University of Maryland.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Veall, M. R. (1991), "Testing for a Global Maximum in an Econometric Context," *Econometrica*, 58, 1959–1965.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 20, 595–601.

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