A Test for Global Maximum

Li Gan† and Jiming Jiang†
University of Texas at Austin and Case Western Reserve University

Abstract

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 that is 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 (1982) 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 over-rejection at relatively small samples.

KEYWORDS: maximum likelihood, multiple roots, consistency, asymptotic efficiency, large sample test, 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 which 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. Since the global maximizer is typically a root to the likelihood equation, the so-called Wald consistency solves, from a theoretical point of view, 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 since the global maximizer is likely (but not for sure) 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 amount of time to find a root, which makes it uneconomical or even impossible to find all the roots. (We have not mentioned the fact that there are cases in which the likelihood equation may have an unbounded number of

† The two authors contribute equally to this work.

roots, 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", i.e., 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 which corresponds to a root to the likelihood equation may still be consistent (e.g., LeCam (1979), see also Lehmann (1983), Example 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), §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 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 way of grid search. It requires a tremendous number of computations, and it often becomes impractical if the support of the parameter space is large and/or the parameter space is multi-dimensional.

In this paper, 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

$$E_{\theta_0}\left(\frac{\partial l}{\partial \theta}|_{\theta_0}\right) = 0 \qquad (1.1)$$

but also

$$E_{\theta_0}\left(\frac{\partial l}{\partial \theta}|_{\theta_0}\right)^2 + E_{\theta_0}\left(\frac{\partial^2 l}{\partial \theta^2}|_{\theta_0}\right) = 0\ , \qquad (1.2)$$

where θ is an unknown parameter and θ_0 is the true θ. Equation (1.1) is consistent with the likelihood equation

$$\frac{\partial l}{\partial \theta} = 0\ . \qquad (1.3)$$

The problem is that there may be multiple roots to (1.3): global maximizer, local maximizer, (local) minimizer, etc. Our claim is very simple: a global maximizer would satisfy

$$\left(\frac{\partial l}{\partial \theta}\right)^2 + \frac{\partial^2 l}{\partial \theta^2} \approx 0 \qquad (1.4)$$

while an inconsistent root (which may correspond to a local maximizer or something else) would not. Note that (1.4) is consistent with (1.2). Therefore this may be rephrased intuitively as that a "good" root is one that makes both (1.1) and (1.2) consistent with their counterparts, (1.3) and (1.4). To demonstrate this numerically, consider the following example.

Example 1.1 (Normal Mixture). A normal mixture density can be written as

$$f(x,\theta) = \frac{p}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{1-p}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right) , \tag{1.5}$$

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 & Makov (1985, §5.5): one corresponds to a local maximum, and the other to a global maximum. Let the values be given as $\sigma_1 = 1$, $\mu_2 = 8$, $\sigma_2 = 4$, $p = 0.4$, and the true value of $\theta = -3$. We consider two ways of computing the standard errors: $\text{SD}_1 = (E_\theta((\partial/\partial\theta)\log f(X,\theta))^2)^{-1/2}$ (the outer product form), and $\text{SD}_2 = (-E_\theta(\partial^2/\partial\theta^2)\log f(X,\theta))^{-1/2}$ (the Hessian form). To see what might happen in practice, a random sample of size 5000 is generated by a computer. The results are shown in the following table.

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	0.0234	0.0492
SD_2	0.0237	0.2807

It is clear that for the global maximum the two ways of computing the SD's give virtually the same result, indicating (1.4); while this is not true for the local maximum.

In §2 we shall make it clear mathematically what "≈ 0" means in (1.4), and prove that the claim is true under reasonable conditions. Based on the result, a simple large sample test for the consistency and asymptotic efficiency of the root is proposed and applied to two examples in §3.

It should be noted that the test used here is that given by White (1982). While the later uses the test to detect model misspecification, we use the same test for another purpose, that is, to detect whether a root corresponds to a global maximum. The basic idea of White's test is the following: 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 (1982), (B) holds as an assumption and (A) is the null hypothesis. In this paper, the roles of (A) and (B) are reversed, i.e., (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 paper, 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 paper. [1]

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 if the model is correct. The answer is that in such a case, if the test rejects, 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 these two. However, in many cases, other techniques are available for model checking. For example, in the case of i.i.d. observations there are methods of 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 *et. al.* (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 therefore searches for a new root, and tests, again, for global maximum. As a contrast, to the best of our knowledge, there is no statistical method 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 paper.

2 Characteristics for consistency and asymptotic efficiency

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

(i) The parameter space Θ is an open interval (not necessarily finite).

(ii) $f(x,\theta) \neq 0$ a.e. μ for all $\theta \in \Theta$.

(iii) $f(x,\theta)$ 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.

[1] In White (1982), when one rejects, implying "possible inconsistency of the QMLE for parameters of interest."

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 shall use $E(\cdot)$ for $E_{\theta_0}(\cdot)$ to simplify notation. We assume that

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

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

$$d_j(\theta) \quad = \quad \int \varphi_j(x,\theta)f(x,\theta_0)d\mu \ , \quad j = 1,2 \ . \tag{2.1}$$

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 $\hat\theta$ be a root to (1.3). The following theorem gives necessary and sufficient condition for the consistency of $\hat\theta$.

Theorem 2.1. Suppose conditions (i)— (iv) are satisfied. In addition, suppose there is $M > 0$ such that

$$\left\{E\left(\inf_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta)\right)\right\} \vee \left\{-E\left(\sup_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta)\right)\right\} \quad > \quad 0 \tag{2.2}$$

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

$$d^{-1}(\{(0,0)\}) \quad = \quad \{\theta_0\} \ . \tag{2.3}$$

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

$$\frac{1}{n}\sum_{i=1}^n \varphi_2(X_i,\hat\theta) \quad \xrightarrow{P} \quad 0 \ . \tag{2.4}$$

Conditions (i) — (iv) are similar to the regularity conditions of Lehmann (1983, page 406): conditions (i) — (iii) corresponds to Lehmann's conditions (i) — (iv), and condition (iv) to Lehmann's condition (vi). Furthermore, condition (2.2) weakens the assumption of White (1982, Assumption 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, (2.2) is satisfied. For the most part, (2.2) says that the parameter space does not have to be compact, but the root to the equation

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

which corresponds to the information identity (1.2), lies asymptotically in a compact subspace. Note that

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

and (2.2) means that either

$$E\left(\inf_{\theta\in\Theta\cap[-M,M]^c} \varphi_2(X_1,\theta)\right) \quad > \quad 0 \ , \tag{2.5}$$

or

$$E\left(\sup_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_1,\theta)\right) \quad < \quad 0\,. \tag{2.6}$$

Intuitively, either (2.5) or (2.6) 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 (2.2) to hold. To see a simple example, consider Example 2.2 in the following. It is shown that $\varphi_2(X_1,\theta) = (X_1 - e^\theta)^2 - e^\theta$. Thus, it is easy to show that (2.5), hence (2.2), holds (see Appendix). Note that in this example the parameter space is not compact, $\Theta = (-\infty,\infty)$.

Perhaps the most notable condition in Theorem 2.1 is (2.3), which is an assumption on the distribution of $X_1,\dots,X_n$, i.e., on $f(x,\theta)$. We now use examples to further illustrate it.

Example 2.1 (Normal Mixture). Consider, again, the normal mixture distribution of Example 1.1. Although there is no analytic expression, the values of $d(\theta)$ can be computed by 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 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, (2.3) is satisfied.

Insert Figure 1 here

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

Example 2.2 (Poisson). Suppose $X_1 \sim \text{Poisson}(e^\theta)$. Then $f(x,\theta) = (e^\theta)^x\exp(-e^\theta)/x! = \exp(\theta x - e^\theta - \log x!)$, $x = 0,1,2,\dots$. 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 $\theta_0 = 0$, $\theta = 0.2$, then $d_2(\theta) \approx -0.1724$. In fact, in this example, $d_2(\theta) = 0$ has two roots: θ_0 and $\log(1 + e^{\theta_0})$, while $d_1(\theta) = 0$ has a single root θ_0. Note that the situation here is on the contrary to that in Example 2.1 in which $d_1(\theta) = 0$ has two roots and $d_2(\theta) = 0$ has one. But the bottom line is: in both cases (2.3) is satisfied.

The next example is selected from an econometrics book.

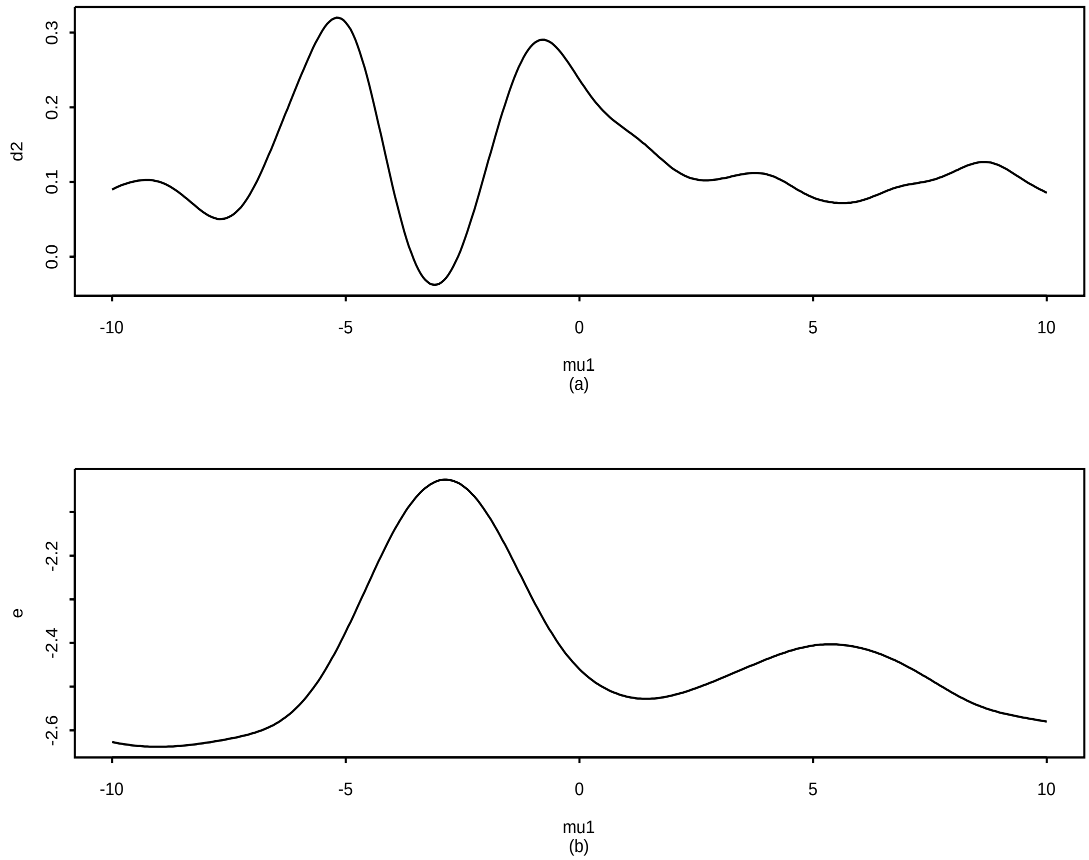

Figure 1: The plots of $d_2(\theta)$ and $e(\theta)$ for the normal mixture with $\theta = \mu_1$ on the x-axis.

Example 2.3 (Amemiya 1994). Suppose $X_1 \sim N(\theta, \theta^2)$. Then, direct calculation shows that

$$\begin{aligned} d_1(\theta) &= -\frac{1}{\theta} - \frac{\theta_0}{\theta^2} + \frac{2\theta_0^2}{\theta^3}\,, \\ d_2(\theta) &= \frac{2}{\theta^2} + \frac{4\theta_0}{\theta^3} - \frac{8\theta_0^2}{\theta^4} - \frac{8\theta_0^3}{\theta^5} + \frac{10\theta_0^4}{\theta^6}\,. \end{aligned}$$

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, (2.3) is, again, satisfied.

After seeing three examples some may conjecture that, perhaps, (2.3) is always satisfied. This turns out, again, to be wrong.

Example 2.4 (A Counterexample). Let $\theta_0 < \theta_1$ be fixed. Let

$$\varphi(\theta) \;=\; \frac{1}{4}\left[1+\frac{\theta-\theta_0}{3(\theta_1-\theta_0)}\right] \,, \quad \psi(\theta) \;=\; \frac{1}{6}\left[2+\frac{(\theta_1-\theta)^3}{(\theta_1-\theta_0)^3}\right] \,.$$

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

Let $a < b < c$ be real numbers. Let the parameter space for θ be (A, B), and 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 (2.3) 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 2.1 to hold, (2.3) cannot be dropped. To see this, let, for simplicity, $\theta_j = j$, $j = 0, 1$. Let $X_1, \ldots, X_n$ be i.i.d. sample from $f(\cdot,\theta_0)$. Then it can be shown that with probability tending to one there is a root $\hat{\theta}_1$ to the likelihood equation such that $\hat{\theta}_1 \xrightarrow{P} \theta_1 \neq \theta_0$, although all the conditions of Theorem 2.1 except (2.3) are satisfied and $(1/n)\sum_{i=1}^n \varphi_2(X_i, \hat{\theta}_1) \xrightarrow{P} 0$ (see Appendix).

Note. Although the counterexample shows that Theorem 2.1 cannot hold without (2.3), it does not necessarily imply limitation on the applicability of the test which we shall propose in the next section. This is because Example 2.4 is rather artificial. In most practical situations, (2.3) is expected to hold. Also note that (2.3) is needed only for the "if" part of Theorem 2.1 and Theorem 2.2 in the following (see the proof in Appendix). Furthermore, the idea of our test for global maximum comes from the observation that the following hold when $\theta = \theta_0$:

$$d_j(\theta) \;\equiv\; E_{\theta_0}\left[\frac{f^{(j)}(X_1,\theta)}{f(X_1,\theta)}\right] \;=\; 0\,, \tag{2.7}$$

$j = 1, 2$, where $f^{(j)}(x,\theta) = (\partial^j/\partial\theta^j)f(x,\theta)$. However, as Example 2.4 shows, there may be $\theta \neq \theta_0$ that satisfies (2.7). This is why condition (2.3) is needed. Note that (2.3) means that

$$d_j(\theta) \;=\; 0\,, \quad j = 1, 2 \quad \text{if and only if} \quad \theta = \theta_0\,. \tag{2.8}$$

More generally, note that, in fact, $\theta = \theta_0$ satisfies a series of equations, i.e., (2.7) with $j = 1, 2, 3, \ldots$ The implication is that, for example, a distribution not satisfying (2.8) may actually satisfy the following:

$$d_j(\theta) \quad = \quad 0 \,, \quad j = 1,2,3 \quad \text{if and only if} \quad \theta = \theta_0 \,. \tag{2.9}$$

For example, it is easy to show that for the distribution in Example 2.4, which breaks (2.8), $d_3(\theta_1) = -3/4(\theta_1 - \theta_0)^3 \neq 0$, therefore (2.9) is satisfied. Thus, one may consider a generalized test which is based on $\varphi_3(X_i, \hat{\theta})$ (defined likewise) instead of $\varphi_2(X_i, \hat{\theta})$, and expect the conclusion of Theorem 2.1 and Theorem 2.2 to hold for a broader class of distributions (i.e., replacing (2.8) by (2.9)). In a related paper, Andrews (1997) proposed a stopping-rule for global minimum in the context of general method of moments assuming there is only one root in the moment conditions. Theorem 2.1 and (2.9) in this paper may be considered as extra moment conditions to ensure such an assumption to hold and his stopping-rule may then be used. [2]

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 (iii). We now assume, in addition to (i)—(iv), the following:

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

(vi) $f(x, \theta)$ 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 2.2. Suppose conditions (i)—(vi) are satisfied, and (2.2) (for some $M > 0$) and (2.3) hold. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \quad \xrightarrow{\mathcal{L}} \quad N(0, I(\theta_0)^{-1}) \tag{2.10}$$

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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_2(X_i, \hat{\theta}) \quad \xrightarrow{\mathcal{L}} \quad N(0, \mathrm{var}(\psi(X_1, \theta_0))) \,. \tag{2.11}$$

In addition to the assumptions explained before, conditions (v) and (vi) are similar to the regularity conditions (v) and (vi) in Lehmann (1983, page 406; also see Theorem 2.3 therein).

An equivalence of the "only if" part of Theorem 2.2 has been obtained by White (1982, Theorem 4.1), which states that under suitable conditions the normalized test statistic, which corresponds to the square of our test

[2]We thank Don Andrews' insights on this point.

statistic in the univariate case, has an asymptotic χ^2 distribution. Despite the fact that the two theorems deal with different problems, an obvious contribution in Theorem 2.2 is that it gives a necessary and sufficient condition for the asymptotic normality, while Theorem 4.1 of White (1982) only provides sufficient conditions.

Also, the conditions of Theorem 2.2 is somehow more general than those of Theorem 4.1 of White (1982). 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 2.2 weakens such a condition to (2.2). As discussed earlier, this allows our theorem to have broader applicability.

The proofs of Theorem 2.1 and Theorem 2.2 are given in 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, then (2.11) implies that

$$\sqrt{n}\phi_2(\hat\theta)/s(\hat\theta) \quad \xrightarrow{\mathcal{L}} \quad N(0,1)\,. \tag{3.1}$$

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. Since $E\varphi_j(X_1, \theta_0) = 0$, $j = 1, 2$, we have

$$\begin{aligned}
(s(\theta_0))^2 &= E(\psi(X_1,\theta_0))^2 \\
&= E(\varphi_2(X_1,\theta_0))^2 + 2\frac{E\phi_2'(\theta_0)}{I(\theta_0)}E\varphi_1(X_1,\theta_0)\varphi_2(X_1,\theta_0) + \left(\frac{E\phi_2'(\theta_0)}{I(\theta_0)}\right)^2 E(\varphi_1(X_1,\theta_0))^2 \\
&= E(\varphi_2(X_1,\theta_0))^2 \\
&\quad + I(\theta_0)^{-1}\left[2E\varphi_1(X_1,\theta_0)\varphi_2(X_1,\theta_0)E\frac{\partial}{\partial\theta}\varphi_2(X_1,\theta)|_{\theta_0} + \left(E\frac{\partial}{\partial\theta}\varphi_2(X_1,\theta)|_{\theta_0}\right)^2\right]\,.
\end{aligned} \tag{3.2}$$

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

In practice, it is sometimes more convenient to use an alternative statistic which 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 Monte Carlo method (i.e., *bootstrap*).

To illustrate the test, we carry out Monte Carlo studies using two previously discussed examples, Example 1.1 and Example 2.3.

Example 1.1 revisited: This example deals with the normal mixture which has been frequently used (e.g., Hamilton (1989)). It is also one of the classical situations where the likelihood equation may have multiple roots. For these reasons we conduct a simulation study on the performance of the large sample test proposed above in the case of Example 1.1.

We consider sample sizes $n = 1000$, 500, and 250. In many economical problems, the data sizes are much larger, thus even $n = 1000$ is not impractical. For each sample size we simulated 500 data sets from a normal mixture distribution with parameters $\mu_1 = -3$, $\mu_2 = 8$, $\sigma_1 = 1$, $\sigma_2 = 4$, and $p = 0.4$. For each data set the test is applied to both the global and local maximizer (of the log-likelihood) found. At significance levels $\alpha = 0.05$ and 0.10, the observed significance level and power at the alternative (i.e., the local maximizer) are reported in Table 2.

It is seen that for all sample sizes the observed significance level is identical or very close to α. For $n = 1000$ or 500 the power is either 1.0 or very close to that. On the other hand, the power drops dramatically in the case of $n = 250$. 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). The simulation is repeated for $n = 500$, and 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 out of 500 cases we are able to find the local maximizer, and hence report the observed power based on those 314.

Table 2: Level and Power for Normal Mixture

	$\mu_2 = 8$						$\mu_2 = 2$	
	$n = 1000$		$n = 500$		$n = 250$		$n = 500$	
	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)
$\alpha = .05$	.05	1.00	.05	.94	.05	.43	.05	.97
$\alpha = .10$	.11	1.00	.13	.99	.11	.65	.13	.99

Example 2.3 revisited: Consider Example 2.3, which comes from Amemiya (1994), $N(\theta, \theta^2)$. Table 3 lists the level and power with 500 replications. We use $\theta_0 = 1$ to generate the data.

Table 3: Level and Power for Amemiya (1994)

	$n = 1000$		$n = 500$		$n = 250$	
	Global (level)	Local (power)	Global (level)	Local (power)	Global (level)	Local (power)
$\alpha = .05$	.064	1.00	0.066	.980	.104	.866
$\alpha = .10$	.098	1.00	0.122	.998	.162	.908

In this case, when the number of observations is relatively large, one has correct level and good power. The test suffers the problem of over-rejection at relatively smaller sample. The small sample property of this test certainly deserves further investigation.

4 Concluding remarks

Although in this paper 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 multi-dimensional.

The property we rely on in this test is the information matrix equality, which holds asymptotically at the global maximum. The size of the test will always be correct asymptotically given a correctly specified model. The asymptotic consistency of the test depends on condition (2.3). A rather artificial counter example is constructed to show that condition (2.3) cannot be dropped universally. However, we also illustrate that a modification of (2.3) based on higher derivatives is potentially useful.

Acknowledgment. The authors wish to thank Professors Peter J. Bickel and Daniel L. McFadden for encouragement and suggestions, and Professors 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 our manuscript and for their insightful comments.

Appendix

That (2.5) holds for Example 2.2. Let $a = \inf_{\theta\in[-M,M]^c} \varphi_2(X_1,\theta)$. Suppose M is suitably large. Then, if $1 \le X_1 \le (1/2)e^M$, we have $\varphi_2(X_1,\theta) \ge (1-e^\theta)^2 - e^\theta > 1 - 3e^{-M}$, $\theta < -M$; and $\varphi_2(X_1,\theta) > (e^\theta/2)^2 - e^\theta \ge (1/4)e^{2M} - e^M$, $\theta > M$. Therefore, $a \ge (1-3e^{-M}) \wedge ((1/4)e^{2M} - e^M)$. 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 $X_1 > (1/2)e^M$. Suppose $\theta > M$. If $\varphi_2(X_1,\theta) \le 0$, then $e^\theta \le X_1 + |X_1 - e^\theta| \le X_1 + e^{\theta/2} \le X_1 + (1/2)e^\theta$. 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 $\theta < -M$. 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 $a \ge -[(4X_1^2 + 2X_1) \vee ((X_1 \vee 1)^2 + 1)]$. It follows that

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

$$-[(4X_1^2+2X_1)\vee((X_1\vee 1)^2+1)]P\left(X_1>\frac{1}{2}e^M\right)$$

$$\longrightarrow \quad P(X_1\geq 1) \;>\; 0\,,$$

as $M\to\infty$. Therefore for suitably large M, (2.5) is satisfied.

Proof of Theorem 2.1. First we note that $d_j(\theta)=E\varphi_j(X_1,\theta)$. WLOG, assume $\Theta\cap[-M,M]^c\neq\emptyset$ and therefore contains a point θ_*. Then (2.2) implies that either $0<\delta_1=E\inf_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_1,\theta)\leq E\varphi_2(X_1,\theta_*)<\infty$ or $0>\delta_2=E\sup_{\theta\in\Theta\cap[-M,M]^c}\varphi_2(X_1,\theta)\geq E\varphi_2(X_1,\theta_*)>-\infty$.

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

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

Therefore (2.4) holds.

Now suppose $P(\hat\theta\in\Theta)\to 1$ and (2.4) holds. We first show that

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

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

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

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

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

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

Suppose that for some $\delta_0>0$ it is not true that $P(|\hat\theta-\theta_0|\geq\delta_0)\to 0$. WLOG, suppose

$$P(|\hat\theta-\theta_0|\geq\delta_0) \;\geq\; \epsilon_0 \tag{A.2}$$

for some $\epsilon_0>0$ and all n. Let $\phi(\theta)=(\phi_1(\theta),\phi_2(\theta))$. Then (iv) implies that $E\phi(\cdot)$ is continuous. Let

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

Since $\theta_0 \notin \Theta_1 = \{\theta \in \Theta : \delta_0 \le |\theta - \theta_0| \le M + |\theta_0|\}$, we have by (2.3) that $\rho = \inf_{\theta \in \Theta_1} |E\phi(\theta)| > 0$. Let $\eta = \rho/4(B_1 \vee B_2)$, then there is 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 $1 \le l \le m$ such that $|\theta_l - \theta| < \eta$. Suppose that $\hat\theta \in \Theta$, $|\hat\theta| \le M$, and $|\hat\theta - \theta_0| \ge \delta_0$. Then $\hat\theta \in \Theta_1$, and hence there is $\tilde\theta \in \{\theta_l, 1 \le l \le m\}$ such that $|\tilde\theta - \hat\theta| < \eta$. Thus by Taylor expansion and SLLN,

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

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

$$\begin{aligned}
P(|\phi_2(\hat\theta)| \ge \rho/4) &\ge P(|\phi_2(\hat\theta)| \ge \rho/2 - o_P(1), |o_P(1)| \le \rho/4) \\
&\ge P(|\phi_2(\hat\theta)| \ge \rho/2 - o_P(1)) - P(|o_P(1)| > \rho/4) \\
&\ge P(\hat\theta \in \Theta, |\hat\theta| \le M, |\hat\theta - \theta_0| \ge \delta_0) - P(|o_P(1)| > \rho/4) \\
&\ge P(|\hat\theta - \theta_0| \ge \delta_0) - P(\hat\theta \notin \Theta) - P(|\hat\theta| > M) - P(|o_P(1)| > \rho/4) \\
&\ge \epsilon_0 - o(1) \quad \longrightarrow \quad \epsilon_0 ,
\end{aligned}$$

as $n \to \infty$, which contradicts (2.4). Therefore we must have $\hat\theta \xrightarrow{P} \theta_0$. □

Proof of Theorem 2.2. Suppose (2.10) holds. Then, by (i), $P(\hat\theta \in \Theta) \to 1$. 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

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

On the other hand, we have

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

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

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

By SLLN, $\phi_j'(\theta_0) \xrightarrow{a.s.} E\phi_j'(\theta_0)$, $j = 1, 2$. Let $S = \{\hat{\theta} \in \Theta, |\hat{\theta} - \theta_0| \leq \delta\}$. Then (i) and (2.10) implies that $P(S) \to 1$. Since on S we have

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

and, by (vi), the right hand side (RHS) 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 (v), imply that the second term on RHS of (A.5) is $o_P(1)$. Similarly, the third term on RHS of (A.5) $= (1/2)\phi_2''(\theta_2) \cdot \sqrt{n}(\hat{\theta}-\theta_0) \cdot (\hat{\theta}-\theta_0) = o_P(1)$. Finally, by the central limit theorem (CLT), the first term on RHS of (A.5) $= (1/\sqrt{n})\sum_{i=1}^n \psi(X_i, \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathrm{var}(\psi(X_1,\theta_0)))$. Therefore (2.11) holds.

Now suppose $P(\hat{\theta} \in \Theta) \to 1$ and (2.11) holds. Then $\phi_2(\hat{\theta}) \xrightarrow{P} 0$, and hence by Theorem 2.1, $\hat{\theta} \xrightarrow{P} \theta_0$. On the other hand, we have by CLT that $\sqrt{n}\phi_1(\theta_0) \xrightarrow{\mathcal{L}} N(0, I(\theta_0))$. (2.10) thus follows by (A.3). $\square$

A proof regarding Example 2.4. Let $I_x = \{1 \leq i \leq n : X_i = x\}$, and $n_x = |I_x|$ ($|\cdot|$ denotes cardinality), $x \in S$. Then it is easy to show that the likelihood equation is equivalent to

$$\hat{s}(\theta) \;=\; \left(\frac{n_a}{n}\right)\frac{\varphi'(\theta)}{\varphi(\theta)} + \left(\frac{n_b}{n}\right)\frac{\psi'(\theta)}{\psi(\theta)} - \left(1 - \frac{n_a}{n} - \frac{n_b}{n}\right)\frac{\varphi'(\theta)+\psi'(\theta)}{1-\varphi(\theta)-\psi(\theta)} \;=\; 0 \, . \tag{A.7}$$

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

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

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

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

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), $\hat{s}(1-\delta) \xrightarrow{a.s.} s(1-\delta)$, $\hat{s}(1+\delta) \xrightarrow{a.s.} s(1+\delta)$, thus

$$P_{\theta_0}(\exists \text{ root to } \hat{s}(\theta) = 0 \text{ in } (1-\delta, 1+\delta)) \;\geq\; P_{\theta_0}\left(\hat{s}(1-\delta) \geq \frac{s(1-\delta)}{2}, \hat{s}(1+\delta) \leq \frac{s(1+\delta)}{2}\right) \;\longrightarrow\; 1 \, . \tag{A.10}$$

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 which is 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 (i)—(iv) and (2.2) 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,

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

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

References

[1] Amemiya, T. (1994), *Introduction to Statistics and Econometrics*, Harvard Univ. Press, Cambridge, Massachusetts.

[2] Andrews, D. (1997), A stopping rule for the computation of generalized method of moments estimation, *Econometrica*, Vol 65, 913-931

[3] Barnett, V. D. (1966), Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots, *Biometrika* 53, 151-166.

[4] Cramér, H. (1946), *Mathematical methods of statistics*, Princeton Univ. Press, Princeton, NJ.

[5] De Haan, L. (1981), Estimation of the minimum of a function using order statistics, *Journal of the American Statistical Association*, 76(374), 467-469.

[6] Hamilton, J. B. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* 57, 357-384.

[7] Jiang, J., Lahiri, P., and Wu, C. (1998), On Pearson-χ^2 testing with unobservable cell frequencies and mixed model diagnostics, submitted.

[8] LeCam, L. (1979), Maximum likelihood: An introduction, *Lecture Notes in Statistics* 18, Univ. of Maryland, College Park, MD.

[9] Lehmann, E. L. (1983), *Theory of Point Estimation*, John Wiley & Sons, New York.

[10] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons.

[11] Veall, M. R. (1991), Testing for a global maximum in an econometric context, *Econometrica*, 58(6), 1959-1965.

[12] Wald, A. (1949), Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* 20, 595-601.

[13] White. H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica* 50, 1-25.

Department of Economics
University of Texas at Austin
Austin, Texas 78712
U. S. A.

Department of Statistics
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-7054
U. S. A.