Testing for a Global Maximum of the Likelihood

Abstract

When several roots to the likelihood equation exist, the root corresponding to the global maximizer of the likelihood is generally retained but this procedure supposes that all possible roots are identified. Since, in many cases, the global maximizer is the only consistent root, we propose a test to detect if a given solution is consistent. This test relies on some necessary and sufficient conditions for consistency of a root that are established. Possibility to employ this test in a badly specified model case is also discussed. Monte-Carlo studies and a real example show that the proposed procedure leads to encouraging results. In particular, it clearly outperforms another available test of this kind, especially for quite small sample sizes.

Keywords. Consistency, maximum likelihood estimator, local and global maximizers, test power, model misspecification.

1 Introduction

In many applications where the maximum likelihood principle is involved, statisticians know that there may be multiple roots to the likelihood equation. Under standard regularity conditions, theory tells us there is a unique consistent root to the likelihood equation (see Cramér 1946 and its multidimensional generalization in Tarone and Gruenhage 1975), but generally gives poor indication on which root is consistent in case of several roots. The review paper of Small et al. (2000) discusses various approaches for selecting among the roots (see also a discussion in Lehmann 1983 chap. 6), including for instance iterating from consistent estimators, employing a bootstrap method or examining the asymptotics when explicit formulas for roots are available. Another possibility is to simply select the root leading to the maximum likelihood value since Wald (1949)
established consistency of the global maximizer of the likelihood under some conditions (typically the global maximizer is a root although it is not always true, in particular for some Gaussian mixtures as noticed first by Kiefer and Wolfowitz 1956). Note also that Wald’s properties of the maximum likelihood estimator (MLE) are generalized by White (1982) in the more realistic case where the probability model is not correctly specified. So, except the rare cases where the global MLE may be inconsistent (see examples in Neyman and Scott 1948 or more recently in Ferguson 1982 or also in Stefanski and Carroll 1987 among others), the strategy which consists of selecting the global maximizer seems to be a straightforward procedure to retain an adequate root. However, some practical difficulties occur and we aim to address them in the present paper.

Indeed, in practice, a search for all roots corresponding to local maximizers may take considerable time and no guarantee is given that all local maximizers will have been found in a finite time, even if the number of roots is bounded (see Barnett 1966 for an example of unbounded number of roots). Beyond this basic strategy of searching, few previous studies are available. For instance, De Haan (1981) proposed a \( p \)-confidence interval of the maximum likelihood value based on extreme-value asymptotic theory. As pointed out by Veall (1991) in an econometric context, this approach becomes impractical because of the number of computations when the support of the parameter space is large and/or the parameter space is multidimensional. In contrast, Markatou et al. (1998) propose a random starting point method based on bootstrap to construct automatically a reasonable search region. Another approach may consist in constructing a test for consistency of a given root to the likelihood equation. In other words, such a method allows to decide if a given root should be adopted as a global maximizer of the likelihood function. Thus, it is possible to search for a new root and to test it until the current root is not rejected. Heyde (1997) and Heyde and Morton (1998) have proposed either to employ a goodness-of-fit criterion to select the best root or to pick the root for which the Hessian of the log-likelihood behaves asymptotically like its expectation evaluated at the root at hand. In the same spirit, Gan and Jiang (1999) (GJ99 in short below) chose a statistic of decision which is based on the difference between the product form of the Fisher expected information matrix about the parameter and its Hessian form. Unfortunately, the Monte-Carlo experiments in the restricted case of a unidimensional parameter highlight a very low power with relatively small sample sizes. As a consequence, it is difficult of recommender the
In this paper, we present a similar procedure to this one of GJ99 to test if a root to the likelihood equation is consistent. Difference is primarily in the employed statistic, which is now simply based on the difference between two different expected log-likelihood expressions. Noting $\ell(\theta)$ the log-likelihood of an unknown parameter whose the true value is $\theta_0$, we know that, under some regularity conditions, $\ell$ satisfies

$$
[E_\theta \nabla \ell(\theta)]_{\theta = \theta_0} = 0,
$$

and obviously

$$
[E_\theta \ell(\theta)]_{\theta = \theta_0} - [E_\theta \ell(\theta)]_{\theta = \theta_0} = 0.
$$

Noting $\hat{\theta}_n$ any root to the likelihood equation

$$
[\nabla \ell(\theta)]_{\theta = \hat{\theta}_n} = 0,
$$

this equation is consistent with (1). The problem is that there may be multiple roots (global maximizer, local maximizer, stationary point and so on). Idea is very simple: A global maximizer would satisfy also

$$
[\ell(\theta)]_{\theta = \hat{\theta}_n} - [E_\theta \ell(\theta)]_{\theta = \hat{\theta}_n} \approx 0,
$$

whereas an inconsistent root (a local maximizer or something else) would not. Indeed, this equation is consistent with (2). So, intuitively, the test would retain as a “good” root a value which verifies both Equations (3) and (4).

Implementing the new method is particularly easy and applicability to multidimensional parameter cases is straightforward. Through experiments, it appears that the power of the proposed test highly outperforms this one of the GJ99’s method. As a consequence, to consider multidimensional parameters situations may be now not completely meaningless. We discuss also possibility to use the procedure with a bad specified model. This last property is useful for practical situations since, often, a model has to be retained through criteria themselves assuming that one optimal likelihood value is already known. It is for instance the case with criteria penalizing the maximum likelihood value as AIC (Akaike 1974) or BIC (Schwarz 1978).
The study is organized as follow. Data, assumptions and theoretical tools to built the test are presented in Section 2. A discussion relating to a badly specified model and also a short presentation of the GJ99’s test are also available. Experiments through simulations and a real data set are then provided in Section 3. In particular, comparison with results of the GJ99’s test is performed. In last section, we conclude this paper with a discussion.

2 Construction of the test

2.1 Two main propositions

The test relies on two propositions that we present following some presentation of data, notations and hypotheses.

Let \( X_1, X_2, \ldots, X_n \) be \( n \) independent random vectors with the same distribution than a variable \( X \) that has density function \( f_t(x) \). Consider also an identifiable parametric density family \( f_t(x; \theta) \) where \( \theta \) is possibly a multidimensional parameter. Define \( \theta_0 \) the value of \( \theta \) where the expected likelihood \( E_t \ln f_t(X; \theta) = \int \ln f_t(x; \theta) f_t(x) dx \) is maximized. In the sequel, we assume that \( \theta_0 \) is unique. Suppose that the following regularity conditions hold:

(a) The parameter space \( \Theta \) is a compact space of which the parameter value \( \theta_0 \) is an interior point.

(b) \( f(x; \theta) \neq 0 \) a.e. for all \( \theta \in \Theta \).

(c) \( f(x; \theta) \) is twice differentiable with respect to \( \theta \) and the integral \( \int f(x; \theta) d\eta \) can be twice differentiated under the integral sign.

Let \( \varphi_1(x; \theta) = \nabla \ln f(x; \theta) \) and \( \varphi_2(x; \theta) = \ln f(x; \theta) - E_\theta \ln f(X; \theta) \). Noting \( | \cdot | \) for the \( L_1 \) norm, we assume the following:

(d) \( \max_{j=1,2} B_j < \infty \) with \( B_j = E_t \sup_{\theta \in \Theta} |\nabla \varphi_j(X; \theta)| \).

(e) \( \max_{j=1,2} E_t \sup_{\theta \in \Theta} |\varphi_j(X; \theta)| < \infty \).

We note \( \phi_j(\theta) = \sum_{i=1}^n \varphi_j(X_i; \theta)/n \) and \( d_j(\theta) = E_t \varphi_j(X; \theta) = E_t \phi_j(\theta) \) \( (j = 1, 2) \). Let define vectors \( \varphi(x; \theta) = (\varphi_1(x; \theta), \varphi_2(x; \theta)) \), \( \phi(\theta) = (\phi_1(\theta), \phi_2(\theta)) \) and \( d(\theta) = (d_1(\theta), d_2(\theta)) \). Let \( \ell(\theta) = \)
\[ \sum_{i=1}^{n} \ln f(X_i; \theta) \] be the log-likelihood function of \( \theta \) based on observations \( X_1, \ldots, X_n \) and let \( \hat{\theta}_n \) be a root to the first derivative of \( \ell(\theta) \), so equivalently a root to \( \phi_1(\theta) \).

Finally we define the variance \( v(\theta) = \text{Var}_{\theta} \ln f(X; \theta) \) and we assume that

\[ (f) \sup_{\theta \in \Theta} |\nabla v(\theta)| < \infty. \]

Now, we present the two propositions which are needed to construct the test. The first one gives necessary and sufficient conditions for convergence of \( \hat{\theta}_n \) towards \( \theta_0 \). The second one gives the asymptotic distribution of \( \phi_2(\hat{\theta}_n) \) under this hypothesis of convergence. Proofs of both propositions are given in Appendix A.

**Proposition 1** Suppose that conditions (a)-(e) are satisfied. In addition, suppose that

\[ d(\theta) = (0, d_2(\theta_0)) \Rightarrow \theta = \theta_0. \]  

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

\[ \phi_2(\hat{\theta}_n) \xrightarrow{P} d_2(\theta_0). \]  

**Proposition 2** Suppose that conditions (a)-(d) are satisfied.

If \( \hat{\theta}_n \xrightarrow{P} \theta_0 \), then

\[ \phi_2(\hat{\theta}_n) \xrightarrow{D} N \left(d_2(\theta_0), \frac{\text{Var}_{\theta_0} \ln f(X; \theta_0)}{n} \right). \]  

### 2.2 A test for convergence

We have now everything to build a test for convergence of \( \hat{\theta}_n \). We distinguish two different situations: the case where the model is correctly specified and the case where it is not. Then, we briefly point out the principle of the GJ99’s test to highlight the differences with the new proposed test.

#### 2.2.1 Case of a correct model

A correct model corresponds to the situation where \( f_1(x) \) belongs to the family \( f(x; \theta) \) and, consequently, \( f_1(x) = f(x; \theta_0) \). In this situation, it is immediate that \( d_2(\theta_0) = 0 \). So, testing for consistency of \( \hat{\theta}_n \) as null hypothesis is equivalent, from Proposition 1, to test for \( \phi_2(\hat{\theta}_n) \xrightarrow{P} 0 \).

Proposition 2 provides the distribution of \( \phi_2(\hat{\theta}_n) \) under the null hypothesis. At this point, an estimator of the variance \( v(\theta_0) = \text{Var}_{\theta_0} \ln f(X; \theta_0) \) is required to perform the test. The two following
estimators are natural: a parametric one \( v(\hat{\theta}_n) = Var_{\hat{\theta}_n} \ln f(X; \hat{\theta}_n) \) where \( \theta_0 \) is simply replaced by \( \hat{\theta}_n \), and a semi-parametric one \( V_n(\hat{\theta}_n) = \sum_{i=1}^{n} (\ln f(X_i; \hat{\theta}_n) - \ell(\hat{\theta}_n)/n)^2/n \) where expectations are estimated in an empirical way. Appendix B discusses properties of these two estimators: both are consistent but the first one, \( v(\hat{\theta}_n) \), has a lower mean squared error value and leads also to better performance of the test. As a consequence, we retain \( v(\hat{\theta}_n) \) as estimator of \( v(\theta_0) \). So, under the null hypothesis, we have

\[
\sqrt{n} \frac{\phi_2(\hat{\theta}_n)}{\sqrt{v(\hat{\theta}_n)}} \overset{D}{\rightarrow} N(0,1). \tag{8}
\]

Note that, in practice, both terms \( E_{\hat{\theta}_n} \ln f(X; \hat{\theta}_n) \) (used in the \( \phi_2(\hat{\theta}_n) \) function) and \( v(\hat{\theta}_n) \) may be easily estimated by a Monte-Carlo method if closed forms are not available.

Here are some comments on this test:

- Note that \( d_2(\theta) \) can be as well positive or negative, so a two-tailed test is required. For instance, take (14) given by a particular family \( f(x; \theta) \) that we will consider later in experiments. It can be shown that if \( \theta_0 > 0 \) and \( \theta > \theta_0 \) then \( d_2(\theta) > 0 \) but if \( \theta < \theta_0 \) then \( d_2(\theta) < 0 \).

- Clearly, Equation (5) is a weak point of this test since it seems difficult to establish some general conditions about its validity. Note that the GJ99’s test is also based on a similar conjecture through Equation (8) in their paper.

However, condition (5) may be expected to hold for many usual classes of density functions \( f_t(x) \) and \( f(x; \theta) \) that appear in most practical situations. For instance, it will be verified for examples we will use in experiments. Note that GJ99 exhibited a counterexample of their own (but similar) conjecture. This counterexample is rather artificial and, consequently, it does not necessarily imply limitation on the applicability of the test. We could hope same properties in our situation.

It is worth noting also that condition (5) is only required to prove convergence of \( \hat{\theta}_n \) in Proposition 1. Thus, the test could be also applied if (5) was not verified since the convergence of \( \hat{\theta}_n \) remains a sufficient condition. Consequences on the test will be: If the null hypothesis is rejected then convergence of \( \hat{\theta}_n \) is rejected too, but if the null hypothesis is preserved nothing could be concluded on convergence of \( \hat{\theta}_n \).
One of the advantages of MLE is that it is invariant both under reparameterization of the model and under a monotone transformation of the sample space. Thus, the procedure which selects the global maximum of the likelihood is fully invariant. While the MLE is invariant (or equivariant) under transformations of the sample space, the log-likelihood is not, even after standardization. Consequently, the proposed test could have the property that a root may pass for one coordinate system and may fail the test for another coordinate system. Nevertheless, for a given significance level, the test will asymptotically gives the same decision independently of the coordinate system.

### 2.2.2 Case of a badly specified model

Consider now that the model is badly specified, it means that densities $f_t(x)$ and $f(x; \theta_0)$ are different: it is the general situation with real data sets. Testing convergence of $\hat{\theta}_n$ towards $\theta_0$ as null hypothesis is still equivalent to test $\phi_2(\hat{\theta}_n) \xrightarrow{P} d_2(\theta_0)$, but now $d_2(\theta_0)$ may be different from zero. Since value $d_2(\theta_0)$ is generally unknown, the test is impracticable.

Nevertheless, if the family $f(x; \theta)$ is flexible enough to allow $f(x; \theta_0)$ “not too far” from $f_t(x)$ in the sense that $|d_2(\theta_0)|$ “small”, the procedure presented in Section 2.2.1 might be used for moderate sample sizes $n$. Of course, when the sample size tends to infinity, the null hypothesis tends to be systematically rejected, since actually $d_2(\theta_0) \neq 0$. But, in many practical situations, the sample size is well-known to be moderate.

In the following, some experiments will illustrate behaviour of the new test in both situations: good model and wrong (but flexible enough) model. We will also compare it with the GJ99’s test in cases available in their paper. So, it is now useful to briefly present their test.

### 2.2.3 Relationship with the GJ99’s test

Difference between GJ99’s test and our proposal is essentially in the formulation of the term $d_2(\theta)$.

Considering that the model is correct (i.e. $f_t(x) = f(x; \theta_0)$), $d_2(\theta)$ is now a matrix (it was a scalar in our proposal) defined by

$$d_2^{GJ}(\theta) = E_{\theta_0}[\nabla \ln f(X; \theta)\nabla \ln f(X; \theta)] + E_{\theta_0}[\nabla^2 \ln f(X; \theta)].$$
So, it is the difference between the outer product form of the Fisher expected information matrix about $\theta$ and its Hessian form. Condition (8) in the GJ99’s paper corresponds to our condition (5).

The null hypothesis consists in testing $\phi_{GJ}^n(\hat{\theta}_n) \xrightarrow{D} 0$ with

$$
\phi_{GJ}^n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ln f(X_i; \hat{\theta}_n) \nabla \ln f(X_i; \hat{\theta}_n)' + \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ln f(X_i; \hat{\theta}_n).
$$

Considering only the unidimensional case, distribution of $\phi_{GJ}^n(\hat{\theta}_n)$ under the null hypothesis is

$$
\frac{\phi_{GJ}^n(\hat{\theta}_n)}{\sqrt{\text{Var}_{\theta_0} \phi_{GJ}^n(\hat{\theta}_n)}} \xrightarrow{D} N(0, 1),
$$

and the variance in the denominator is approximated by a Monte-Carlo method.

3 Experiments

3.1 A simple mixture case

We consider the normal mixture distribution of Example 1 in GJ99, that is

$$
f(x; \theta) = p\psi(x; \mu_1, \sigma_1^2) + (1-p)\psi(x; \mu_2, \sigma_2^2),
$$

where $\theta = \mu_1$ and $\psi(x; \mu, \sigma^2)$ is the univariate normal density of mean $\mu$ and variance $\sigma^2$, and $p$ ($0 < p < 1$) is the mixing proportion of the first component. The likelihood equation for $\theta$ has typically two roots if the two normal means are “well-separated” (e.g. Titterington, Smith and Makov 1985); one corresponds to a local maximizer and the other to a global maximizer. Values are the following: $\sigma_1^2 = 1$, $\sigma_2^2 = 16$ and $p = 0.4$ and the true value of $\theta = -3$.

Because there is no analytic expression, the values $E_{\theta_0} \ln f(X; \theta)$ and $d_2(\theta)$ are computed by a Monte-Carlo method. Figure 1 shows that the global maximizer of the expected likelihood is $\theta_0 = -3$ and the local maximizer is somewhere between 5 and 10. It appears that only the global maximizer leads to $d_2(\theta) = 0$ and, consequently, condition (5) is verified.

[Figure 1 about here.]

We consider different sample sizes $n$: 5, 10, 50, 100, 250, 500, 1000. For each sample size, we simulated 500 datasets from the true distribution. For each dataset, we applied the test to both the global and the local maximizer of $\ell(\theta)$ found. Figure 2 reports the observed significance level
and the observed power at the alternative (i.e. the local maximizer) at significance levels $\alpha = 0.05$ and 0.10. Note that GJ99’s results are also displayed for the available values of $n$ in their article: 100, 250 and 1000. We note a high improvement of the power with the new test.

[Figure 2 about here.]

3.2 A particular normal distribution

We consider now Example 4 of GJ99 that was employed before in an econometrics context by Amemiya (1994). It corresponds to the normal distribution $N(\mu; \mu^2)$. Direct calculation shows that

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

and

$$d_2(\theta) = -\frac{\theta_0}{\theta^2} (\theta_0 - \theta).$$

It is easy to show that $d_1(\theta)$ has two roots: $\theta_0$ and $-2\theta_0$. However, $d_2(-2\theta_0) = -\frac{3}{4} < 0$ and so condition (5) is verified for this example too.

[Figure 3 about here.]

Figure 3 displays the levels and the power with 500 replications in the case $\theta_0 = 1$. Note that experimental conditions are the same than in the previous example. We notice that power is higher with the new test again. But, surprisingly, the empirical level is very low in comparison with the theoretical significance level. We can explain this result by noting that, under the null hypothesis, the variance of $\phi_2(\hat{\theta}_n)$ is equal to $\text{Var}_{\theta_0}[X/\hat{\theta}_n]/4$ (see Proposition 6 in Appendix C) and that this value is less than its asymptotic variance $\nu(\hat{\theta}_n)$ that is equal to $1/(2n)$ (see Proposition 5 in Appendix C). Figure 4 clearly demonstrates this fact.

[Figure 4 about here.]

**Remark** In the classical multinormal situation $N(\mu, \Sigma)$, Proposition 7 (Appendix C) shows that $\phi_2(\hat{\theta}_n) = 0$ and, so, the true variance of $\phi_2(\hat{\theta}_n)$ is null and obviously less than the asymptotic variance $1/(2n)$. Of course applying the test in the classical normal case has no interest since only
one root to the likelihood equation exists, but we note it would lead to a test with significance
level equal to zero.

3.3 Some multi parameter examples

3.3.1 A two parameter case

Consider the previous normal mixture case with the two unknown centers $\mu_1$ and $\mu_2$, so $\theta = (\mu_1, \mu_2)$
and the true parameter is $\theta_0 = (-3, 8)$. In Figure 5, $E_{\theta_0} \ln f(X; \theta)$ and $d_2(\theta)$ are displayed and we
note that two maximizers exist (a local and a global one) and, moreover, condition (5) is verified
since only the global maximizer leads to $d_2(\theta) = 0$.

[Figure 5 about here.]

[Figure 6 about here.]

With different sample sizes ($n = 10, 25, 50, 100$), and the same other experimental conditions
than in the simple mixture case, Figure 6 reports the observed significance level and power at
significance levels $\alpha = 0.05$ and 0.10. The power is still reasonably good but, not surprisingly, is
lower that in the one parameter case. In order to measure influence of increasing the number of
free parameters on the power, let us consider now a situation with more parameters to estimate.

3.3.2 A ten parameter case

We retain a bivariate normal mixture with five components and same mixing proportions. The
first four components are defined as follow: their centers are on the nodes of a square with side
6 and variances matrices are equal to identity. The center of the fifth component corresponds to
the center of the square and its variance matrix is four times the identity matrix. Figure 7 (a)
provides a sample from this model with $n = 500$.

Only centers have to be estimated, so ten parameters are unknown. Because of the symmetry,
two different maxima of the value of the expected likelihood exist: a global and a local one. Note
that it is difficult to verify condition (5) and so we do not. Figure 7 (b) displays the levels and power
with 500 replications, different sample sizes ($n = 50, 100, 500, 1000$) and two different significance
levels ($\alpha = 0.05, 0.10$). Clearly, we note that the power is now low for moderate sample sizes.
3.4 Case of a badly specified model

We consider now a situation where the true density \( f_t(x) \) does not belong to the family \( f(x; \theta) \). We choose \( f_t(x) \) as being the following Gaussian mixture with three components:

\[
f_t(x) = 0.4\psi(x; -3, 1) + 0.3\psi(x; 5, 9) + 0.3\psi(x; 11, 9).
\] (15)

The model \( f(x; \theta) \) is the same as (12), so \( \theta = \mu_1 \): It is a mixture of two Gaussian components with only one free center. It is usual to have a bad number of components specification in many model-based mixture contexts (see for instance McLachlan and Peel 2000, Chap. 6). Figure 8 exhibits difference between the two densities \( f_t(x) \) and \( f(x; \theta_0) \), \( \theta_0 \) being the value of \( \theta \) maximizing the expected likelihood \( E_t \ln f(X; \theta) \). Values of this likelihood and also \( d_2(\theta) \) are displayed in Figure 9. We note that the global maximizer is \( \theta_0 = -3 \) and that two local maximizers exist: one somewhere between 3 and 6 and the other somewhere between 10 and 13. It appears that the global maximizer does not lead to \( d_2(\theta_0) = 0 \) but, clearly, \( d_2(\theta_0) \) is “not too far” from zero in comparison to the values of \( d_2 \) obtained with the two local maximizers. Note also that (5) is verified again since \( d_2(\theta_0) \neq d_2(\theta) \) for any value \( \theta \neq \theta_0 \).

Figure 10 displays the observed level and power for both local maximizers with 500 replications of \( f_t(x) \) for different sample sizes \((n = 5, 10, 25, 50, 100, 250, 500, 1000)\) and significance level \( \alpha = 0.05 \). Results are similar for both local maxima. Note that the observed power is quickly close to one and, as expected, observed level monotonically increases with \( n \). When \( n \) is relatively small (says between 10 and 100), power is high and level is sufficiently low to justify using the test.
3.5 A real data set

We consider now the Old Faithful data (the version from Venables and Ripley 1994) which consists of data on 272 eruptions of the Old Faithful geyser in Yellowstone National Park. Each observation is composed by two measurements: the duration (in minutes) of the eruption and the waiting time (in minutes) before the next eruption.

We retain a bivariate normal mixture model with three components but with equal proportions and equal variance matrices between components. Estimation of the 9 free parameters of the model is performed with the EM algorithm (Dempster et al. 1977), and implementation of this particular model at the E step of EM is given, among other models, in Celeux and Govaert (1995). The algorithm is run 100 times at random for 1000 iterations and we obtain only two different solutions presented in Figure 11. Next, the test of convergence is applied to both solutions of the likelihood and the two corresponding P-values are displayed in Table 1. We note a strong evidence for choosing the maximum likelihood solution whereas the local maximum is clearly rejected at any classical significance levels.

Consider now a model with free variance matrices between the three components. This model leads to 15 free parameters. Figure 12 displays the three found solutions of the log-likelihood (other experimental conditions are unchanged) and Table 2 provides corresponding P-values. The lowest local maximum is clearly rejected but the two other solutions (included the maximum likelihood value) show strong evidence. So, we confirm a fact already noted in some previous experiments: the power may be low when the number of free parameters is high comparing to the sample size.

[Figure 11 about here.]

[Table 1 about here.]

[Figure 12 about here.]

[Table 2 about here.]
4 Discussion

In case of multiple roots to the likelihood equation, a standard procedure is to select the root corresponding to the global maximizer of the likelihood. Nevertheless, one is seldom certain to have enumerated all possible roots. Since, in many situations, the MLE is the only consistent root to the likelihood equation, we proposed a test for consistency of any root to the likelihood equation. This test seems quite simple and rather natural. A previous test for a global maximum of the likelihood was already suggested by GJ99 but this test was presented in the restricted univariate parameter case and also led to very low power for moderate sample sizes. As a consequence, investigation towards multivariate parameters situation was not considered by these authors.

Results provided through experiments of the new proposed test in this work seem showing that these difficulties are partially overcomed: Power of the test is highly improved in univariate parameter cases and a bivariate parameter case is successfully considered. Note that the test is particularly straightforward to implement for any dimension. Nevertheless, power of the test could become quite low when the dimension of the parameter space significantly increases in comparison to the sample size. That was highlighted by a ten parameters case and by a real data set.

Extension to a misspecified model is also considered. When the model is “flexible enough” (it means not “too far” from the true distribution), we suggest that the test could be used for moderate sample sizes. Experiments, of which the used real data set, confirms this possibility. It is an important issue since many criteria to select a model rely on knowledge of the maximum likelihood estimator. For instance, it is the case for criteria penalizing the maximum likelihood as AIC (Akaike 1974) or BIC (Schwarz 1978) among others (see McLachlan and Peel 2000 for a review of some other criteria). Nevertheless, it is obvious that if the “flexibility” hypothesis is uncertain, rejection in the test has the two following meanings without possibility to decide between both: Either the consistent root is not reached, or the model is not correct.

Finally, a theoretical aspect of the test relies on condition (5). This one seems often verified as illustrated by experimental situations, but no guarantee is given for other, but usual, density families. Nevertheless, as discussed before in the paper, the test could be still applied if, unfortunately, condition (5) was not true. Note that GJ99’s test relies also on a similar conjecture.
A Proofs of the two main propositions

Proof of Proposition 1 This proof is strongly inspired from Theorem 1 of the GJ99’s paper.

Suppose that \( \hat{\theta}_n \xrightarrow{P} \theta_0 \). By condition (a), we have \( P(\hat{\theta}_n \in \Theta) \to 1 \). Then, Taylor expansion gives:

\[
\phi_2(\hat{\theta}_n) = \phi_2(\theta_0) + \left( \nabla \phi_2(\theta_0) \right)'(\hat{\theta}_n - \theta_0).
\]  

(16)

By the strong law of large number (SLLN), we have \( \phi_2(\theta_0) \overset{a.s.}{\rightarrow} d_2(\theta_0) \) and \( \nabla \phi_2(\theta_0) \overset{a.s.}{\rightarrow} \nabla d_2(\theta_0) \).

When \( \hat{\theta}_n \in \Theta \) and by condition (d) we know that \( |\nabla d_2(\theta_0)| = |E_1 \nabla \varphi_2(X; \theta_0)| \leq E_1 \sup_{\theta \in \Theta} |\nabla \varphi_2(X; \theta)| < \infty \), therefore (6) holds.

Now consider that \( P(\hat{\theta}_n \in \Theta) \to 1 \) and (6) hold. We suppose that \( \hat{\theta}_n \xrightarrow{P} \theta_0 \) is false. So, there are \( \delta_0 > 0 \) and \( \varepsilon_0 > 0 \) such that, for all \( n \),

\[
P(|\hat{\theta}_n - \theta_0| \geq \delta_0) \geq \varepsilon_0.
\]  

(17)

We define \( \Theta_1 = \{ \theta : |\theta - \theta_0| \geq \delta_0 \} \). Since \( d(\cdot) = E_1 \phi(\cdot) \) is continuous (Proposition 4 in Appendix C) and \( \theta_0 \notin \Theta_1 \), we have by (5) that \( \rho = \inf_{\theta \in \Theta_1} |d(\theta) - d(\theta_0)| > 0 \). Let \( \eta = \rho/\|X(1) \cup B_2)\| \). 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 \( |\theta_l \to \theta| < \eta \). Suppose that \( \hat{\theta}_n \in \Theta \) and \( |\hat{\theta}_n - \theta_0| \geq \delta_0 \). Then \( \hat{\theta}_n \in \Theta_1 \) and hence there is \( \bar{\theta} \in \{ \theta_l \} \) such that \( |\bar{\theta} - \hat{\theta}_n| < \eta \). Now, we seek onioration of \( |\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \):

\[
|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq |\phi(\bar{\theta}) - d(\theta_0)| - |\phi(\bar{\theta}) - \phi(\hat{\theta}_n)|
\]  

(19)

\[
\geq \min_{1 \leq l \leq m} |\phi(\theta_l) - d(\theta_0)| - |\phi(\theta_l) - \phi(\hat{\theta}_n)| \geq |\phi(\theta_l) - d(\theta_0)| - |\phi(\hat{\theta}) - \phi(\hat{\theta}_n)|
\]  

(20)

In addition, we have

\[
\min_{1 \leq l \leq m} |\phi(\theta_l) - d(\theta_0)| \geq \min_{1 \leq l \leq m} |d(\theta_l) - d(\theta_0)| - \max_{1 \leq l \leq m} |\phi(\theta_l) - d(\theta_l)|,
\]  

(21)

where

\[
\min_{1 \leq l \leq m} |d(\theta_l) - d(\theta_0)| = \frac{1}{2} \min_{1 \leq l \leq m} |d(\theta_l) - d(\theta_0)| + \frac{1}{2} \min_{1 \leq l \leq m} |d(\theta_l) - d(\theta_0)| \]

(22)

\[
\geq \frac{\rho}{2} + \frac{\rho}{2}
\]  

(23)

\[
\geq \frac{\rho}{2} + \frac{\rho}{2} \frac{B_1 + B_2}{2(B_1 \cup B_2)}
\]  

(24)

\[
= \frac{\rho}{2} + \eta(B_1 + B_2).
\]  

(25)
Moreover, Taylor inequality and the fact that \(|\theta_i - \theta| < \eta\) give (with \(j = 1, 2\))

\[
|\phi_j(\hat{\theta}) - \phi_j(\hat{\theta}_n)| \leq \left| \hat{\theta} - \hat{\theta}_n \right| \left( \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in \Theta} |\nabla \phi_j(X; \theta)| \right) 
\]

\[
\leq \eta \left( \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in \Theta} |\nabla \phi_j(X; \theta)| \right). 
\]

(26)

(27)

Combining Equations (20), (21), (25) and (27), we obtain:

\[
|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \frac{\rho}{2} - \max_{1 \leq i \leq n} |\phi(\theta_i) - d(\theta_i)| - \eta \sum_{j=1}^{2} \left( \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in \Theta} |\nabla \phi_j(X; \theta)| - B_j \right). 
\]

(28)

Then the SLLN gives (recall that \(d(\cdot) = E_1\phi(\cdot)\)):

\[
|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \frac{\rho}{2} - o_P(1). 
\]

(29)

Therefore, with the same \(o_P(1)\), we have

\[
P(|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \rho/4) \geq P(|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \rho/2 - o_P(1), |o_P(1)| \leq \rho/4) 
\]

\[
\geq P(|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \rho/2 - o_P(1)) - P(|o_P(1)| > \rho/4), 
\]

(30)

since for any events \(A\) and \(B\), we know that \(P(A, B) \geq P(A) - P(B^c)\), \(B^c\) being the complement of \(B\). Moreover, Inequality (29) being the consequence of both \(\hat{\theta}_n \in \Theta\) and \(|\hat{\theta}_n - \theta_0| \geq \delta_0\), and for all events \(A \subset B \Rightarrow P(A) \leq P(B)\), we deduce that

\[
P(|\phi_2(\hat{\theta}_n) - d_2(\theta_0)| \geq \rho/4) \geq P(\hat{\theta}_n \in \Theta, |\hat{\theta}_n - \theta_0| \geq \delta_0) - P(|o_P(1)| > \rho/4) 
\]

\[
\geq P(|\hat{\theta}_n - \theta_0| \geq \delta_0) - P(\hat{\theta}_n \notin \Theta) - P(|o_P(1)| > \rho/4) 
\]

\[
\geq \varepsilon_0 - o(1) \to \varepsilon_0 
\]

(32)

(33)

(34)

as \(n \to \infty\), which contradicts (6). Therefore, we must have \(\hat{\theta}_n \xrightarrow{P} \theta_0\).

**Proof of Proposition 2** Suppose that \(\hat{\theta}_n \xrightarrow{P} \theta_0\). Recall Taylor expansion in (16). The central limit theorem leads to

\[
\phi_2(\theta_0) \xrightarrow{D} N \left( d_2(\theta_0), \frac{\text{Var}_\theta \ln f(X; \theta_0)}{n} \right). 
\]

(35)

Moreover, we proved in Proposition 1 that both \(P(\hat{\theta}_n \in \Theta) \to 1\), and, when \(\hat{\theta}_n \in \Theta\), \(\nabla \phi_2(\theta_0)'(\hat{\theta}_n - \theta_0) = o_P(1)\). Then, using Slutsky theorem, (7) holds.
B Selecting an estimator of the variance

The problem is to study behavior of two possible estimators of the variance $v(\theta_0) = \text{Var}_{\theta_0} \ln f(X; \theta_0)$ in the case of a correct model: $v(\hat{\theta}_n) = \text{Var}_{\theta_0} \ln f(X; \hat{\theta}_n)$ and $V_n(\hat{\theta}_n) = \sum_{i=1}^n (\ln f(x_i; \hat{\theta}_n) - \ell(\hat{\theta}_n)/n)^2/n$. First, we show consistency of both. Then, we compare them through the mean squared error criterion and also performance of the test.

B.1 Consistency of both estimators

Proposition 3 Suppose conditions (a), (b) and (f) are satisfied. If $\hat{\theta}_n \xrightarrow{P} \theta_0$, then $v(\hat{\theta}_n) \xrightarrow{P} v(\theta_0)$ and $V_n(\hat{\theta}_n) \xrightarrow{P} V_n(\theta_0)$.

Proof Taylor expansion of $v(\hat{\theta}_n)$ gives

$$v(\hat{\theta}_n) = v(\theta_0) + \left(\nabla v(\theta_0)' + o_P(1)\right)(\hat{\theta}_n - \theta_0).$$

(36)

When $\theta \in \Theta$, condition (f) and convergence $\hat{\theta}_n \xrightarrow{P} \theta_0$ involve that $v(\hat{\theta}_n) \xrightarrow{P} v(\theta_0)$.

In the same manner, Taylor expansion of $V_n(\hat{\theta}_n)$ is

$$V_n(\hat{\theta}_n) = V_n(\theta_0) + \left(\nabla V_n(\theta_0)' + o_P(1)\right)(\hat{\theta}_n - \theta_0).$$

(37)

By SLLN, we have $V_n(\theta_0) \xrightarrow{a.s.} v(\theta_0)$ and $\nabla V_n(\theta_0) \xrightarrow{a.s.} \nabla v(\theta_0)$. Consequently, when $\theta \in \Theta$, $V_n(\hat{\theta}_n) \xrightarrow{P} v(\theta_0)$.

B.2 Empirical comparison of both estimators

We propose to compare by simulation the two candidates with (i) the mean squared criterion (mse) and (ii) their performance on the test. Remind that mse of an estimator $H$ for a real value $h$ is defined by

$$\text{mse}_h[H] = \text{Var}_h[H] + (E_h[H] - h)^2.$$  

(38)

We consider the simple mixture case (Section 3.1) with the same experimental conditions and $\hat{\theta}_n$ the consistent estimator of $\theta_0 = -3$. Figure 13 displays mse and power of the test at significance level $\alpha = 0.05$. Clearly, mse is lower for $v(\hat{\theta}_n)$ and it leads also to a better power for small sample sizes.
C Other propositions and proofs

Proposition 4 If conditions (a), (b), (c) and (e) are satisfied, then $E_t\phi(\cdot)$ is continuous.

Proof For any sequence of parameters $(\theta_n)$ in $\Theta$ such that $\theta_n \to \theta$ as $n \to \infty$ with $\theta \in \Theta$, consider the corresponding sequence of functions $(\phi(x; \theta_n))$. From condition (c), $\phi(x; \cdot)$ is continuous and, so, $\phi(x; \theta_n) \to \phi(x; \theta)$ pointwise. Moreover, condition (e) gives

$$E_t \sup_n |\phi(X; \theta_n)| \leq E_t \sup_n |\phi_1(X; \theta_n)| + E_t \sup_n |\phi_2(X; \theta_n)| < \infty. \quad (39)$$

Dominated convergence theorem implies that $E_t\phi(X; \theta_n) \to E_t\phi(X; \theta)$. Remind that $E_t\phi(X; \cdot) = E_t\phi(\cdot)$ to conclude that $E_t\phi(\cdot)$ is continuous for any $\theta \in \Theta$.

Proposition 5 If $X$ has normal density $\phi(x; \mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$ then

$$E_{\mu, \sigma^2} \ln \phi(X; \mu, \sigma^2) = -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} - \frac{1}{2} \ln(2\pi) \quad (40)$$

and

$$Var_{\mu, \sigma^2} \ln \phi(X; \mu, \sigma^2) = \frac{1}{2}. \quad (41)$$

Proof From $\ln \phi(X; \mu, \sigma^2) = -\ln(\sigma^2)/2 - (X - \mu)^2/(2\sigma^2) - \ln(2\pi)/2$, (40) is obvious and (41) is proven by the following:

$$Var_{\mu, \sigma^2} \ln \phi(X; \mu, \sigma^2) = \frac{Var_{\mu, \sigma^2}(X - \mu)^2}{4\sigma^4} \quad (42)$$

$$= \frac{E_{\mu, \sigma^2}(X - \mu)^4 - E_{\mu, \sigma^2}^2(X - \mu)^2}{4\sigma^4} \quad (43)$$

$$= \frac{3\sigma^4 - \sigma^4}{4\sigma^4} = \frac{1}{2}. \quad (44)$$

Proposition 6 Let $f(x; \theta) = \phi(x; \theta, \theta^2)$, $X_1, \ldots, X_n \text{i.i.d.} f(x; \theta_0)$ and $\bar{X} = \sum_{i=1}^n X_i/n$. If $\hat{\theta}_n$ is a solution of the likelihood equation, then

$$Var_{\theta_0, \phi_2}(\hat{\theta}_n) = \frac{1}{4} Var_{\theta_0} \left[ \bar{X} \right]. \quad (45)$$
Proof} Direct calculation shows that

\[
\frac{\ell(\theta)}{n} = -\frac{1}{2} \ln \theta^2 + \frac{\bar{X}}{\theta} - \frac{\bar{X}}{2\theta^2} - \frac{1}{2} - \frac{1}{2} \ln(2\pi)
\]  

(46)

with \(\bar{X} = \frac{\sum_{i=1}^{n} X_i^2}{n}\) and so, using (40) in Proposition 5, we obtain

\[
\phi_2(\hat{\theta}_n) = \frac{\bar{X}}{\hat{\theta}_n} - \frac{\bar{X}}{2\hat{\theta}_n^2}.
\]  

(47)

Maximizing \(\ell(\theta)\) leads to two roots

\[
\hat{\theta}_n^{(1)} = \frac{1}{2} \left( -\bar{X} + \sqrt{\bar{X}^2 + 4\bar{X}} \right) \quad \text{and} \quad \hat{\theta}_n^{(2)} = \frac{1}{2} \left( -\bar{X} - \sqrt{\bar{X}^2 + 4\bar{X}} \right),
\]  

(48)

so, in both cases we have \(\hat{\theta}_n^2 + \hat{\theta}_n - \bar{X} = 0\) and we deduce that

\[
\phi_2(\hat{\theta}_n) = \frac{\bar{X}}{2\hat{\theta}_n} - \frac{1}{2}.
\]  

(49)

As a consequence, we have \(\text{Var}_{0,\phi_2}(\hat{\theta}_n) = \text{Var}_{0,\phi_2}(\bar{X}/\hat{\theta}_n)/4\).

Proposition 7 Let \(f(x; \theta)\) corresponding to the multinormal density of mean \(\mu\) and variance \(\Sigma\), where \(\theta = (\mu, \Sigma)\). With \(\hat{\theta}_n\) the (unique) solution of the likelihood equation, we have \(\phi_2(\hat{\theta}_n) = 0\).

Proof} We know that \(\hat{\theta}_n = (\hat{\mu}_n, \hat{\Sigma}_n)\) with \(\hat{\mu}_n = \frac{\sum_{i=1}^{n} X_i}{n}\) and \(\hat{\Sigma}_n = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^\prime}{n}\).

Moreover, if \(Y \sim f(x; \hat{\theta}_n)\), it is obvious that \(\hat{\mu}_n = E_{\hat{\theta}_n}Y\) and \(\hat{\Sigma}_n = E_{\hat{\theta}_n}[(Y - \hat{\mu}_n)(Y - \hat{\mu}_n)^\prime]\).

First, we have

\[
\frac{\ell(\hat{\theta}_n)}{n} = -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{1}{2n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^\prime \hat{\Sigma}_n^{-1} (X_i - \hat{\mu}_n) + \text{cst}
\]  

(50)

\[
= -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{1}{2n} \text{tr} \left\{ \sum_{i=1}^{n} (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^\prime \hat{\Sigma}_n^{-1} \right\} + \text{cst}
\]  

(51)

\[
= -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{d}{2} + \text{cst}.
\]  

(52)

Then, we have also:

\[
E_{\hat{\theta}_n} \ln f(Y; \hat{\theta}_n) = E_{\hat{\theta}_n} \left[ -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{1}{2} (Y - \hat{\mu}_n)^\prime \hat{\Sigma}_n^{-1} (Y - \hat{\mu}_n) + \text{cst} \right]
\]  

(53)

\[
= -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{1}{2} \text{tr} \left\{ E_{\hat{\theta}_n}[(Y - \hat{\mu}_n)(Y - \hat{\mu}_n)^\prime] \hat{\Sigma}_n^{-1} \right\} + \text{cst}
\]  

(54)

\[
= -\frac{1}{2} \ln |\hat{\Sigma}_n| - \frac{d}{2} + \text{cst}.
\]  

(55)

We conclude by noting that, by definition, \(\phi_2(\hat{\theta}_n) = \ell(\hat{\theta}_n)/n - E_{\hat{\theta}_n} \ln f(Y; \hat{\theta}_n)\).
References


Figure 1: Plots of (a) $E_{	heta_0} \ln f(X; \theta)$ and (b) $d_2(\theta)$ for the simple mixture case.
Example 1 from Gan & Jiang – Significance level: \( \alpha = 0.05 \) and \( \alpha = 0.10 \).

Figure 2: Level and power for the simple mixture case when (a) \( \alpha = 0.05 \) and (b) \( \alpha = 0.10 \).
Figure 3: Level and power for the particular normal distribution when (a) $\alpha = 0.05$ and (b) $\alpha = 0.10$. 
Figure 4: True variance and asymptotic variance of $\phi_2(\hat{\theta}_n)$ in the particular normal situation.
Figure 5: Plots of (a) $E_{\theta_0} \ln f(X; \theta)$ and (b) $d_2(\theta)$ for the two parameter case.
Example 1 from Gan & Jiang with 2 free centers

Figure 6: Level and power for the two parameter case with $\alpha = 0.05$ and $\alpha = 0.10$. 
Figure 7: The ten parameter case: (a) a sample with isodensity curves and (b) level and power for $\alpha = 0.05$ and $\alpha = 0.10$. 
Figure 8: True density $f_t(x)$ and the optimal density $f(x; \theta_0)$ of the model in a bad specified model case.
Figure 9: Plots of (a) $E_t \ln f(X; \theta)$ and (b) $d_2(\theta)$ for the bad specified model case.
Figure 10: Level and power for the bad specified model case with $\alpha = 0.05$. 
Figure 11: Component ellipses obtained with the Old Faithful geyser for each maxima (equal variance matrices model): (a) $\ell(\hat{\theta}_n)/n = -4.1584$ and (b) $\ell(\hat{\theta}_n)/n = -4.3868$. 
Figure 12: Component ellipses obtained with the Old Faithful geyser for each maxima (free variance matrices model): (a) $\ell(\hat{\theta}_n)/n = -4.1215$, (b) $\ell(\hat{\theta}_n)/n = -4.1439$ and (c) $\ell(\hat{\theta}_n)/n = -4.2991$. 
Figure 13: Comparison of estimators $v(\hat{\theta}_n)$ and $V_n(\hat{\theta}_n)$ in the simple mixture case respectively with: (a) mse criterion and (b) power of the test.
Table 1: P-values for both likelihood solutions of the Old Faithful geyser (equal variance matrices model).

| \( t(\theta_n)/n \) | -4.1584 | -4.3868 |
| \( \phi_2(\theta_n) \) | -0.0062 | -0.1937 |
| P-value            | 0.9153  | 0.0022  |
Table 2: P-values for each likelihood solution of the Old Faithful geyser (free variance matrices model).

<table>
<thead>
<tr>
<th></th>
<th>$\ell(\hat{\theta}_n)/n$</th>
<th>(\phi_2(\hat{\theta}_n))</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.1215</td>
<td>0.0133</td>
<td>0.8358</td>
</tr>
<tr>
<td></td>
<td>-4.1439</td>
<td>-0.0021</td>
<td>0.9713</td>
</tr>
<tr>
<td></td>
<td>-4.2991</td>
<td>-0.3942</td>
<td>0.0000</td>
</tr>
</tbody>
</table>