Testing for a Global Maximum of the Likelihood

Christophe Biernacki

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 and simply consists of comparing the difference between two expected log-likelihood expressions. Monte-Carlo studies and a real life example show that the proposed procedure leads to encouraging results. In particular, it clearly outperforms another available test of this kind, especially for relatively small sample sizes.

KEY WORDS: Consistency; Maximum likelihood; Local and global maximizers; Test power.

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 that 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

Christophe Biernacki is Assistant professor, Department of Mathematics, Université de Franche-Comté, 16 route de Gray, 25030 Besançon Cedex, France (E-mail: biernac@math.univ-fcomte.fr).

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 misspecified. So, except in the rare cases where the 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 an 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 look 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 to recommend the use of this test, especially in the usual situation of multidimensional parameter estimation.

In this paper, we present a procedure similar to GJ99's test in order to decide if a root to the likelihood equation is consistent. The difference primarily lies in the employed statistic, which is now simply based on the difference between two expected log-likelihood expressions. Denoting by $\ell(\theta)$ the log-likelihood of a parameter θ whose true value (unknown) is θ_0, it may exist some values θ, different from θ_0, which satisfy also

$$E_{\theta_0} \nabla \ell(\theta) = 0. \tag{1}$$

Thus, the problem is that there may be multiple roots to the likelihood equation (global maximizer, local maximizer, stationary point and so on), it means multiple roots $\hat{\theta}_n$ such that

$$\left[\nabla \ell(\theta)\right]_{\theta=\hat{\theta}_n} = 0. \tag{2}$$

In order to detect the root $\hat{\theta}_n$ corresponding to a global maximizer of $\ell(\theta)$, the idea of this paper is very simple: A global maximizer would satisfy not only (2), but also

$$\left[\ell(\theta)\right]_{\theta=\hat{\theta}_n} - \left[E_\theta \ell(\theta)\right]_{\theta=\hat{\theta}_n} \approx 0, \tag{3}$$

whereas an inconsistent root (a local maximizer or something else) would not. Indeed, it seems natural that both terms in the left hand side of (3) are two different consistent estimators of the same term $\left[E_\theta \ell(\theta)\right]_{\theta=\theta_0}$ if $\hat{\theta}_n$ is consistent. So, intuitively, the test retains as a "good" root a parameter which verifies both Equations (2) and (3). 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 GJ99's method. As a consequence, to consider multidimensional parameters situations may be now far from meaningless.

The study is organized as follow. Data, assumptions and theoretical tools to built the test are presented in Section 2 where a short presentation of GJ99's test is also available. Simulation experiments and a real data set are then provided in Section 3 to evaluate the performance of the new test. In particular, comparisons with results of GJ99's test are given. In the last section, we conclude this paper with a discussion.

2. CONSTRUCTION OF THE TEST

2.1 Two theorems

Let $X_1, X_2, \ldots, X_n$ be n independent random vectors with the same distribution as a variable X having density function $f_t(x)$. Consider also an identifiable parametric density family $f(x;\theta)$ where θ is possibly a multidimensional parameter. Define θ_0 the value of θ where the expected log-likelihood $E_t \ln f(X;\theta) = \int \ln f(x;\theta) f_t(x) dx$ is maximized. In the sequel, we assume that θ_0 is unique and that the following regularity conditions hold:

(a) The parameter space Θ is a compact space of which the parameter value θ_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 θ and the integral $\int f(x;\theta) d\eta$ is twice differentiable under the integral sign.

First, let us define $\varphi_1(x;\theta) = \nabla \ln f(x;\theta)$ and $\varphi_2(x;\theta) = \ln f(x;\theta) - E_\theta \ln f(X;\theta)$. Set also $\phi_j(\theta) = \sum_{i=1}^n \varphi_j(X_i;\theta)/n$, $d_j(\theta) = E_t \varphi_j(X;\theta) = E_t \phi_j(\theta)$ $(j = 1, 2)$, $\varphi(x;\theta) = (\varphi_1(x;\theta), \varphi_2(x;\theta))$, $\phi(\theta) = (\phi_1(\theta), \phi_2(\theta))$ 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 θ based on observations $X_1, \ldots, X_n$ and let $\hat{\theta}_n$ be a root of $\nabla \ell(\theta)$, so equivalently a root of $\phi_1(\theta)$. Finally, defining the variance $v(\theta) = Var_\theta \ln f(X;\theta)$ and denoting by $|\cdot|$ the L_1 norm, we assume the following conditions:

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

(e) $0 < Var_t \ln f(X;\theta_0) < \infty$.

(f) $|\nabla v(\theta_0)| < \infty$.

We present now two theorems on which our test relies in an essential way. The first one gives necessary and sufficient conditions for convergence of $\hat{\theta}_n$ towards θ_0. The second one gives the asymptotic distribution of $\phi_2(\hat{\theta}_n)$ under this hypothesis of convergence. Proofs of both theorems are given in Appendix A.

Theorem 1. Suppose that conditions (a)-(d) are satisfied. In addition, suppose that

$$d(\theta) = (0, d_2(\theta_0)) \Rightarrow \theta = \theta_0. \tag{4}$$

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

$$\phi_2(\hat{\theta}_n) \xrightarrow{P} d_2(\theta_0). \tag{5}$$

Theorem 2. Suppose that conditions (a)-(e) are satisfied. If $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\phi_2(\hat{\theta}_n) \xrightarrow{D} N\left(d_2(\theta_0), \frac{Var_t \ln f(X;\theta_0)}{n}\right). \tag{6}$$

2.2 A test for convergence

We have now everything to build a procedure for testing convergence of $\hat{\theta}_n$ towards θ_0. We consider the situation where $f_t(x)$ belongs to the family $f(x;\theta)$ and, consequently, $f_t(x) = f(x;\theta_0)$. In this situation, it is immediate that $d_2(\theta_0) = 0$ and, so, testing for consistency of $\hat{\theta}_n$ as null hypothesis is equivalent, from Theorem 1, to test for $\phi_2(\hat{\theta}_n) \xrightarrow{P} 0$. Then Theorem 2 provides the asymptotic distribution of $\phi_2(\hat{\theta}_n)$ under the null hypothesis. At this point, an estimator of the variance $v(\theta_0) = Var_{\theta_0} \ln f(X;\theta_0)$ 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 θ_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 some empirical observations show that the first one, $v(\hat{\theta}_n)$, has a lower mean squared error value and leads also to a better performance of the test. It is reasonable to conjecture that $v(\hat{\theta}_n)$ uses more efficiently the information about the model than $V_n(\hat{\theta}_n)$ does but it seems difficult to derive some general properties to compare more accurately $v(\hat{\theta}_n)$ and $V_n(\hat{\theta}_n)$. Thus, we choose $v(\hat{\theta}_n)$ as an estimator of $v(\theta_0)$ in our test and, under the null hypothesis, we have

$$\sqrt{n}\frac{\phi_2(\hat{\theta}_n)}{\sqrt{v(\hat{\theta}_n)}} \xrightarrow{D} N(0,1). \tag{7}$$

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. Thus, in this work, a i.i.d. sample $Y_1, \ldots, Y_m$ is generated from $f(x;\hat{\theta}_n)$ and both terms are respectively estimated by $A_m = \sum_{i=1}^{m} \ln f(Y_i;\hat{\theta}_n)/m$ and $B_m = \sum_{i=1}^{m} (\ln f(Y_i;\hat{\theta}_n) - A_m)^2/m$.

Here are some comments on this test:

1. Note that $d_2(\theta)$ can be as well positive or negative, so a two-sided test is required. For instance, taking (12) given by a particular family $f(x;\theta)$ considered later in experiments, it can be shown that, when $\theta_0 > 0$, if $\theta > \theta_0$ then $d_2(\theta) > 0$ but if $\theta < \theta_0$ then $d_2(\theta) < 0$.

2. Clearly, Equation (4) is a weak point of this test since it seems difficult to establish some general conditions about its validity. However, condition (4) 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 a similar conjecture needed for their test. This counterexample is rather artificial and, consequently, it does not necessarily imply limitation on the applicability of the test. We can hope for similar properties in our context. In addition, it is worth noting that condition (4) is only required to prove convergence of $\hat{\theta}_n$ in Theorem 1. Thus, the test could be also applied if (4) was not verified since the convergence of $\hat{\theta}_n$ remains a sufficient condition. Consequences on the test will be the following: 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 $\hat{\theta}_n$.

3. 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 the test may fail for another coordinate system. Nevertheless, given a significance level, the test will asymptotically provide the same decision independently of the coordinate system.

4. If the model is misspecified, it means that densities $f_t(x)$ and $f(x;\theta_0)$ are different, testing convergence of $\hat{\theta}_n$ towards θ_0 as the 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 not valid anymore. Nevertheless, we will carry out later some experiments to study the behaviour of our test in such a situation.

2.3 Relationship with 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, $d_2(\theta)$ is now a *matrix* defined by

$$d_2^{GJ}(\theta) = E_{\theta_0}[\nabla \ln f(X;\theta) \nabla \ln f(X;\theta)'] + E_{\theta_0}[\nabla^2 \ln f(X;\theta)]. \tag{8}$$

So, it is the difference between the *outer product form* of the Fisher expected information matrix about θ and its *Hessian form*. Condition (8) in GJ99's paper corresponds to our condition (4). The null hypothesis consists in testing $\phi_2^{GJ}(\hat{\theta}_n) \xrightarrow{P} 0$ with

$$\phi_2^{GJ}(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ln f(X_i;\hat{\theta}_n) \nabla \ln f(X_i;\hat{\theta}_n)' + \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ln f(X_i;\hat{\theta}_n). \tag{9}$$

Considering only the unidimensional case, distribution of $\phi_2^{GJ}(\hat{\theta}_n)$ under the null hypothesis is

$$\frac{\phi_2^{GJ}(\hat{\theta}_n)}{\sqrt{Var_{\theta_0}\phi_2^{GJ}(\hat{\theta}_n)}} \xrightarrow{D} N(0,1), \tag{10}$$

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

$$f(x;\theta) = p\psi(x;\mu_1,\sigma_1^2) + (1-p)\psi(x;\mu_2,\sigma_2^2), \tag{11}$$

where $\theta = \mu_1$ and $\psi(x;\mu,\sigma^2)$ is the univariate normal density of mean μ and variance σ^2, and p $(0 < p < 1)$ is the mixing proportion of the first component. The likelihood equation for θ has typically two roots if the two normal means are "well-separated" (e.g. Titterington *et al.* 1985): one corresponds to a local maximizer and the other to a global maximizer. Values are the following: $\sigma_1^2 = 1$, $\mu_2 = 8$, $\sigma_2^2 = 16$ 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 log-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 (4) is verified. We consider different sample sizes n: 5, 10, 50, 100, 250, 500, 1000. For each sample size, we simulated 500 datasets from $f(x;\theta_0)$ and, for each dataset, we applied the test to both the global and the local maximizer of $\ell(\theta)$. 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. GJ99's results are also displayed for the available values of n in their article ($n = 250$, 500, 1000) and we note a high improvement of the power with the new test.

[Figure 1 about here.]

[Figure 2 about here.]

3.2 A particular normal distribution

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

$$d_1(\theta) = -\frac{1}{\theta} - \frac{\theta_0}{\theta^2} + \frac{2\theta_0^2}{\theta^3} \qquad \text{and} \qquad d_2(\theta) = -\frac{\theta_0}{\theta^2}(\theta_0 - \theta). \tag{12}$$

It is easy to show that $d_1(\theta)$ has two roots: θ_0 and $-2\theta_0$. However, $d_2(-2\theta_0) = -\frac{3}{4} < 0$ and so condition (4) is verified for this example too.

[Figure 3 about here.]

Figure 3 of the appendixes displays the level and the power with 500 replications in the case $\theta_0 = 1$. Note that other experimental conditions are the same than in the previous example. We notice that the 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 $Var_{\theta_0}\left\{\left[-1 + \sqrt{1 + 4n\sum_{i=1}^n X_i^2/(\sum_{i=1}^n X_i)^2}\right]^{-1}\right\}$ (see Proposition 4 in Appendix C) and that this value is less than its asymptotic variance $v(\hat{\theta}_n)/n$ that is equal to $1/(2n)$ (see Proposition 3 in Appendix C). Figure 4 illustrates this fact.

[Figure 4 about here.]

Remark In the normal case $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$, Proposition 5 (Appendix C) shows that $\phi_2(\hat{\theta}_n) = 0$. So, $Var_{\theta_0}\phi_2(\hat{\theta}_n)$ is null and obviously less than the asymptotic variance $1/(2n)$. Of course applying the test in this case has no interest since only one root to the likelihood equation exists, but we note that it would lead to a test with an empirical significance level equal to zero.

3.3 Some multi parameter examples

3.3.1 A two parameter case

Consider the normal mixture of Section 3.1 where centers μ_1 and μ_2 are unknown, 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 (4) is verified since only the global maximizer leads to $d_2(\theta) = 0$. With different sample sizes (n = 10, 25, 50, 100), and the same other experimental conditions as in Section 3.1, 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.

[Figure 5 about here.]

[Figure 6 about here.]

3.3.2 A ten parameter case

We choose to study a bivariate normal mixture with five components and same mixing proportions. The first four components have centers on the nodes of a square with side 6 and variances matrices equal to identity. The center of the fifth component corresponds to the center of the square, which is fixed to $(0, 0)$, and its variance matrix is four times the identity matrix. Figure 7 (a) provides a sample from this model with $n = 500$. Since only centers have to be estimated, only ten parameters are unknown. Because of the symmetry, two different maxima of the expected log-likelihood exist: a global one and a local one which corresponds to the situation where the fifth component is exchanged with one of the four others. Note that it is difficult to verify condition (4) because of the high dimension and so we do not. Figure 7 (b) displays the level and the power with 500 replications, different sample sizes ($n = 50, 100, 500, 1000$) and two different significance

levels ($\alpha = 0.05, 0.10$). In these experiments, the global maximizer is obtained by starting the EM algorithm (Dempster *et al.* 1977) with the true centers whereas a local maximizer is obtained by starting EM after exchanging centers of components 1 and 5. Moreover, EM is stopped after 1000 iterations. Clearly, we note that the power is now low for these sample sizes.

[Figure 7 about here.]

3.4 Case of a misspecified 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:

$$f_t(x) = 0.4\psi(x;-3,1) + 0.3\psi(x;5;9) + 0.3\psi(x;11,9), \tag{13}$$

whereas the model $f(x;\theta)$ is the same as (11), i.e. a mixture of two Gaussian components with only one free center ($\theta = \mu_1$). 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 $f(x;\theta_0)$, θ_0 being the value of θ maximizing the expected log-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 close to $\theta_0 = -3$ and that two local maximizers exist: one somewhere between 3 and 6 (local solution 1) and the other somewhere between 10 and 13 (local solution 2). 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 (4) is verified again since $d_2(\theta_0) \neq d_2(\theta)$ for any value $\theta \neq \theta_0$.

[Figure 8 about here.]

[Figure 9 about here.]

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. The observed power is quickly close to one and, as expected, the observed level monotically increases with n. When n is relatively small (e.g., between 10 and 100), the power is high and the level is sufficiently low to justify the test.

[Figure 10 about here.]

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 for components. Estimation of the 9 free parameters of the model is performed with the EM algorithm, 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 for 1000 iterations from a random starting mixture parameter where proportions are equal, centers are randomly drawn without replication in the dataset and the common variance matrix is equal to λI, λ being uniformly drawn in $[0, 10]$ (see McLachlan and Peel 2000, chap. 2, for a review of some strategies for choosing starting values). Finally, 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.

We consider now a model where the three variance matrices are free, so, this model has 15 unknown parameters. Figure 12 displays the three found solutions of the log-likelihood (other experimental conditions are unchanged except that now the starting variance matrix of the kth component is equal to $\lambda_k I$ where λ_k is uniformly drawn in $[0, 10]$) and Table 2 provides corresponding P-values. The lowest local maximum is clearly rejected but the two other solutions (included the MLE) 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 compared 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 test introduced in this work seem to show that these difficulties are partially overcome: The power of the test is highly improved in univariate parameter cases and a bivariate parameter case is successfully treated. In addition, the test is particularly straightforward to implement in any dimensions. Nevertheless, its power 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 parameter case and on a real life data set.

It is worth noting that the test is theoretically impracticable in the case of a misspecified model although an experiment explores the possibility of employing it in a situation where the true distribution and the model are different but the value $|d_2(\theta_0)|$ is "relatively small". Strictly speaking, rejection in the test has the two following meanings without possibility to decide between both as it was noted before by GJ99 in their context: Either the consistent root is not reached, or the model is not correct. Such a property is problematic in practice since many criteria proposed to select a model such as BIC (Schwarz 1978) among others (see McLachlan and Peel 2000 for a review of some other criteria) rely on the knowledge of the maximum likelihood estimator.

Finally, a theoretical aspect of the test relies on condition (4). This one seems often verified as illustrated by experimental situations, but no guarantee is given for other, albeit usual, density families. Although, as discussed before in the paper, the test could be still applied if, unfortunately,

condition (4) were not true, there is a need to explore more widely its overall validity.

REFERENCES

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

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

Celeux, G., and Govaert, G. (1995), "Gaussian Parsimonious Clustering Models," *Pattern Recognition*, **28**, 781–793.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society B*, **39**, 1–38.

Ferguson, T. S. (1982), "An inconsistent Maximum Likelihood Estimate", *Journal of the American Statistical Association*, **77**, 380, 831–834.

Gan, L., and Jiang, J. (1999), "A Test for Global Maximum," *Journal of the American Statistical Association*, **94**, 447, 847–854.

Heyde, C. C. (1997), *Quasi-Likelihood and Its Application*, New-York: Springer.

Heyde, C. C., and Morton, R. (1998), "Multiple roots in general estimating equations," *Biometrika*, **85**, 954–959.

Kiefer, J., and Wolfowitz, J. (1956), "Consistency of the maximum-likelihood estimation in the presence of infinitely many incidental parameters," *Annals of Mathematical Statistics*, **27**, 887–906.

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

Markatou, M., Basu, A., and Lindsay, B. G. (1998), "Weighted Likelihood Equations with Bootstrap Root Search," *Journal of the American Statistical Association*, **93**, 442, 740–750.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New-York: Wiley.

Neyman, J., and Scott, E. (1948), "Consistent Estimators Based on Partially Consistent Observations," *Econometrica*, **16**, 1–32.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 461–464.

Small, C. G., Wang, J., and Yang, Z. (2000), "Eliminating Multiple Root Problems in Estimation," *Statistical Science*, **15**, 4, 313–341.

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional scores and optimal scores for generalized linear measurement-error models," *Biometrika*, **74**, 4, 703–716.

Tarone, R. D., and Gruenhage, G. (1975), "A note on the uniqueness of roots of the likelihood equations for vector-valued parameters," *Journal of the American Statistical Association*, **70**, 903–904.

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*, **56**, 1959–1965.

Venables, W. N., and Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus*, New-York: Springer-Verlag.

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

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

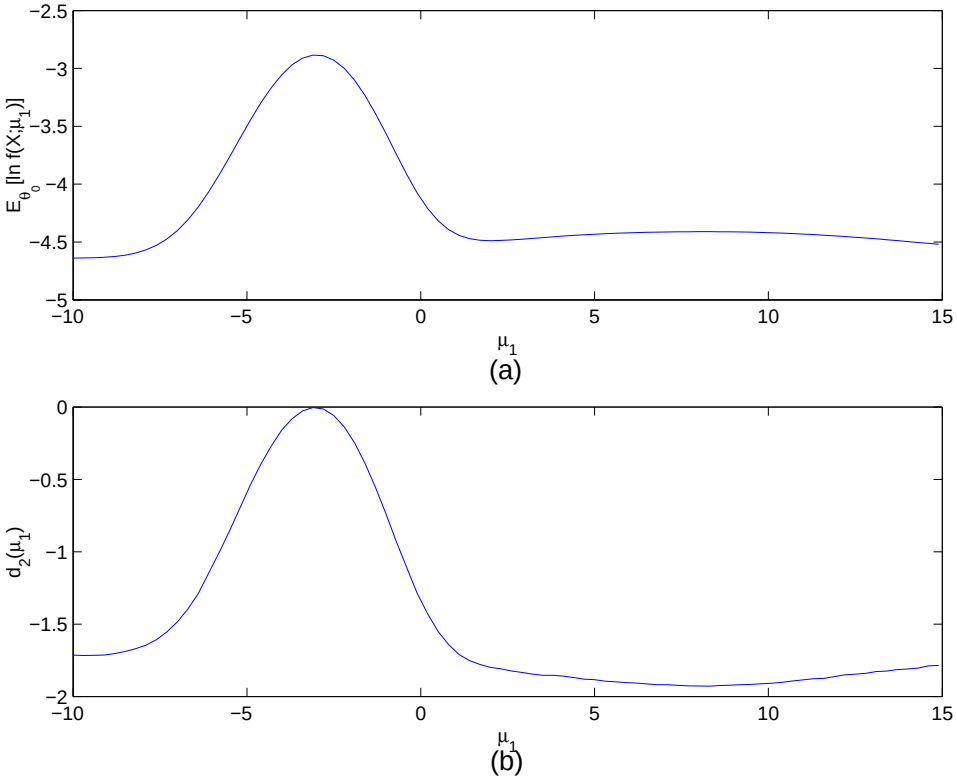
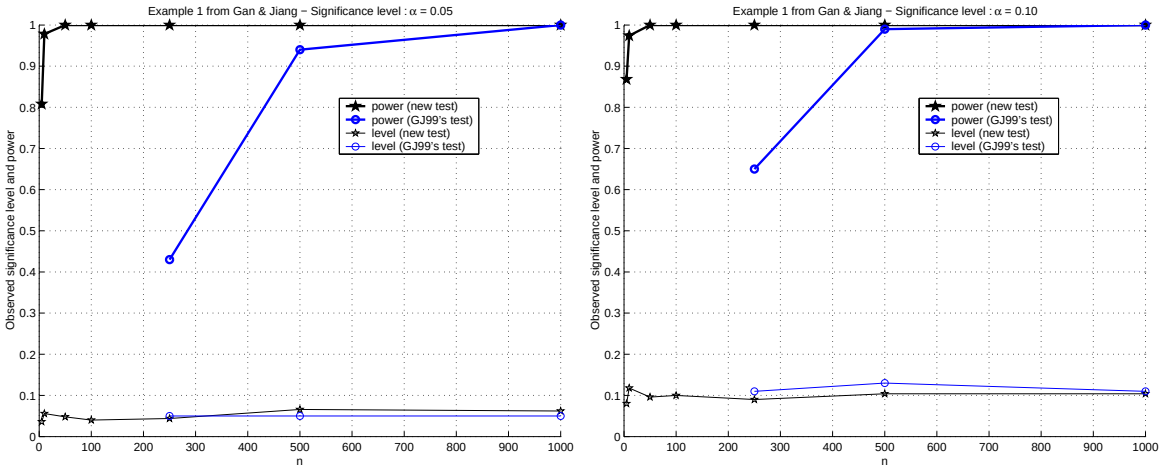
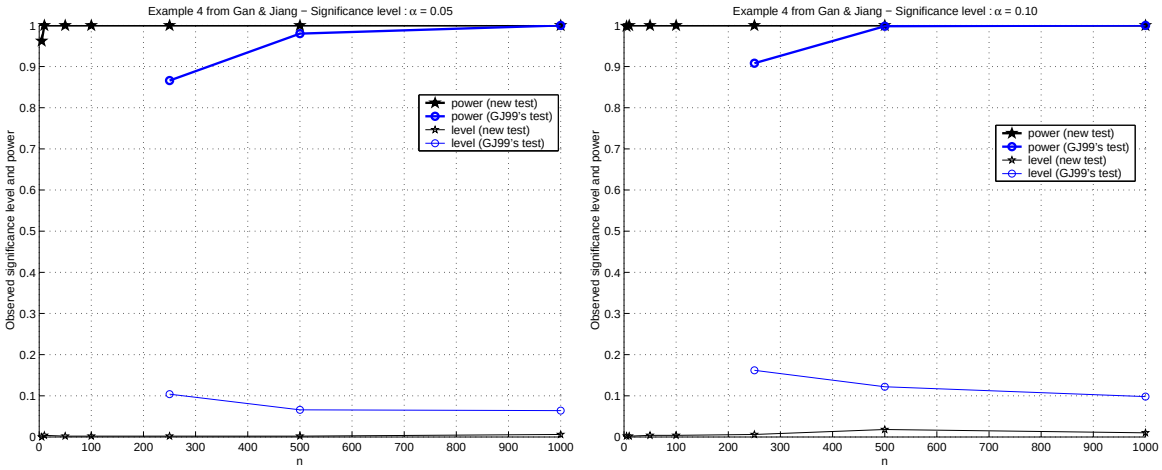


Figure 1. Plots of (a) $E_{\theta_0}[\ln f(X;\theta)]$ and (b) $d_2(\theta)$ for the simple mixture case.



(a) (b)

Figure 2. Level and power for the simple mixture case when (a) $\alpha = 0.05$ *and (b)* $\alpha = 0.10$.



(a) (b)

Figure 3. Level and power for the particular normal distribution when (a) $\alpha = 0.05$ *and (b)* $\alpha = 0.10$.

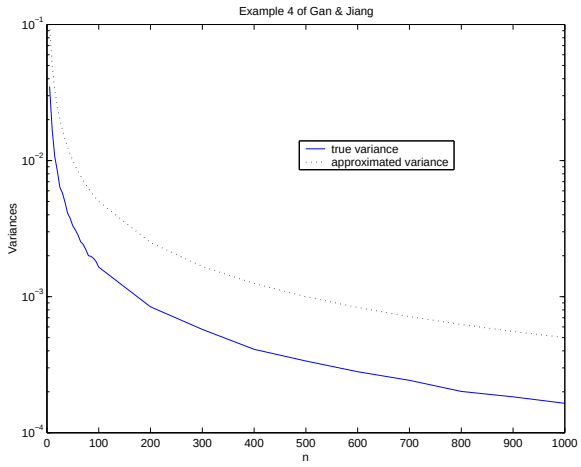


Figure 4. True variance and asymptotic variance of $\phi_2(\hat{\theta}_n)$ in the particular normal situation.

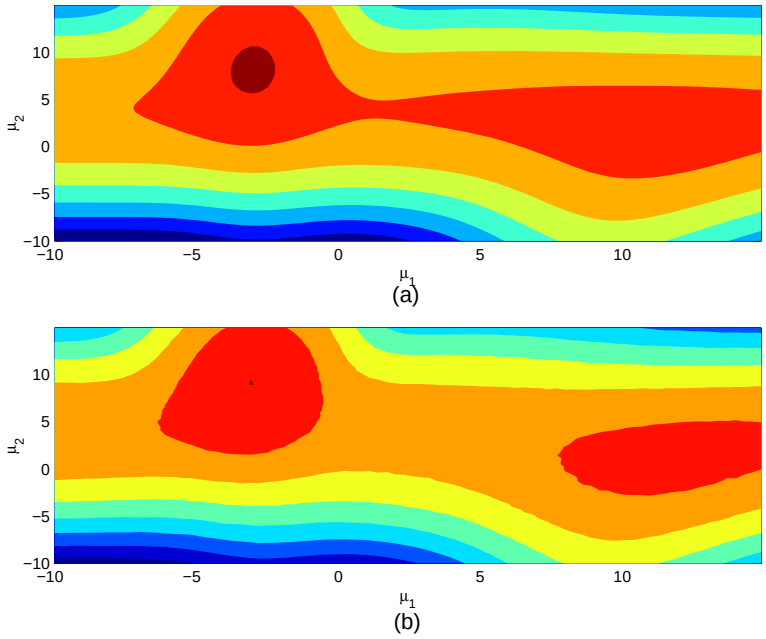


Figure 5. Plots of (a) $E_{\theta_0} \ln f(X;\theta)$ and (b) $d_2(\theta)$ for the two parameter case.

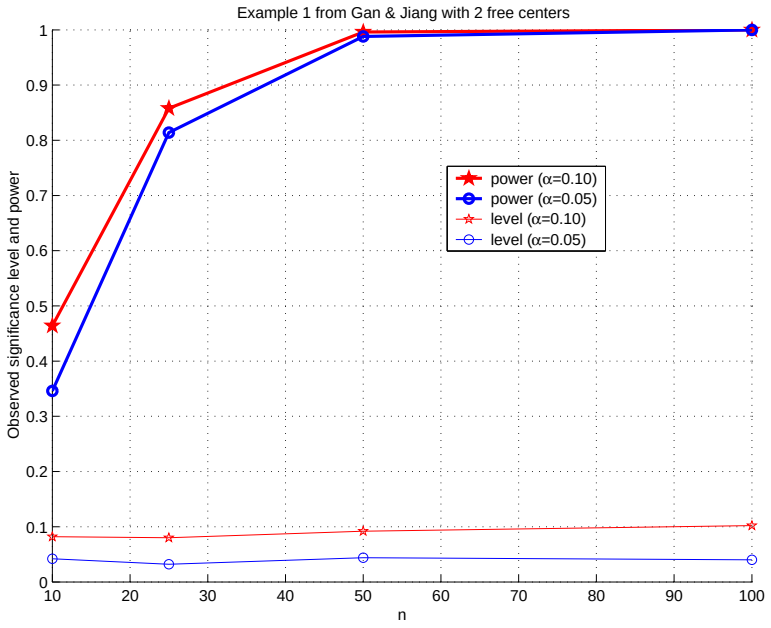
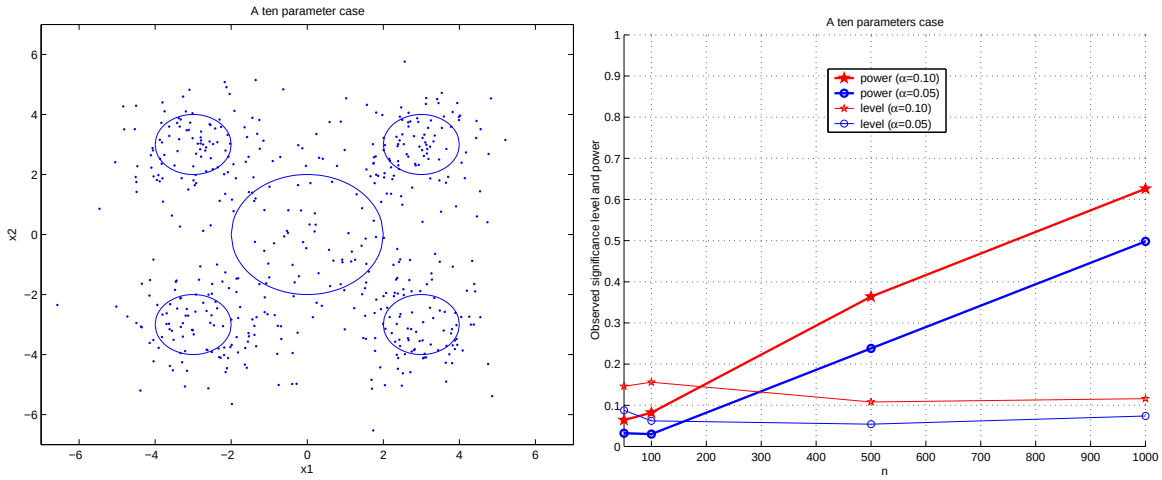


Figure 6. Level and power for the two parameter case with $\alpha = 0.05$ *and* $\alpha = 0.10$.



(a) (b)

Figure 7. The ten parameter case: (a) a sample with isodensity curves and (b) level and power for $\alpha = 0.05$ *and* $\alpha = 0.10$.

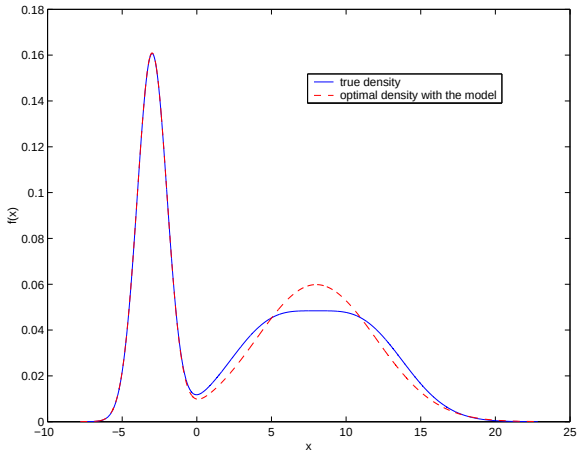


Figure 8. True density $f_t(x)$ and the optimal density $f(x;\theta_0)$ of the model in a misspecified model case.

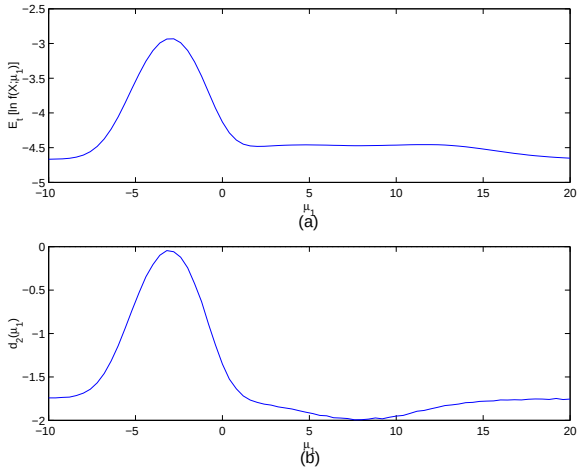


Figure 9. Plots of (a) $E_t \ln f(X;\theta)$ and (b) $d_2(\theta)$ for the misspecified model case.

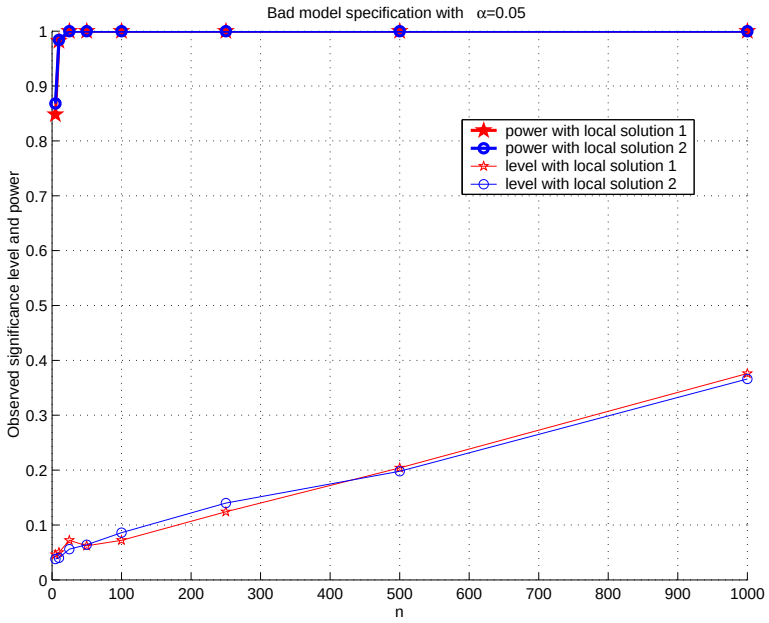
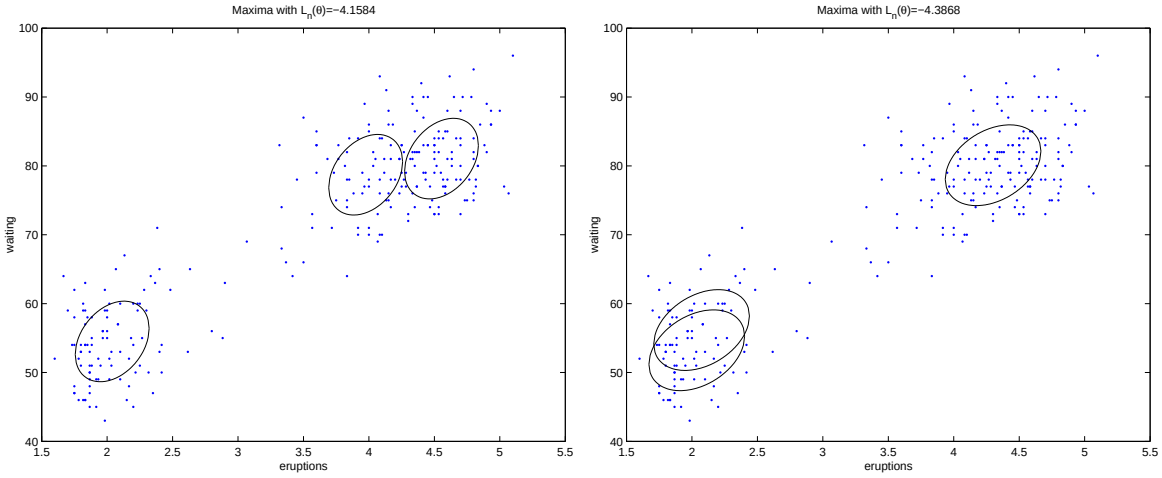
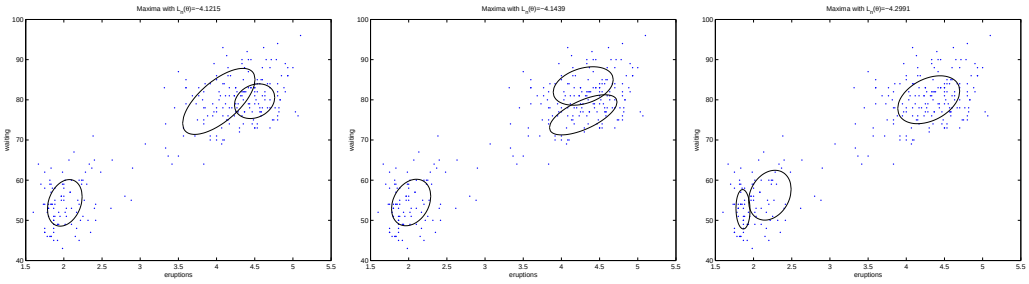


Figure 10. Level and power for the misspecified model case with $\alpha = 0.05$.



(a) (b)

*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)* $\ell(\hat{\theta}_n)/n = -4.3868$.



(a) (b) (c)

*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)* $\ell(\hat{\theta}_n)/n = -4.2991$*.*

$\ell(\hat{\theta}_n)/n$	-4.1584	-4.3868
$\phi_2(\hat{\theta}_n)$	-0.0062	-0.1937
P-value	0.9153	0.0022

*Table 1. P-values for both likelihood solutions of the Old Faithful geyser (*equal *variance matrices model).*

$\ell(\hat{\theta}_n)/n$	-4.1215	-4.1439	-4.2991
$\phi_2(\hat{\theta}_n)$	0.0133	-0.0021	-0.3942
P-value	0.8358	0.9713	0.0000

*Table 2. P-values for each likelihood solution of the Old Faithful geyser (*free *variance matrices model).*