TESTS FOR GLOBAL MAXIMUM OF THE LIKELIHOOD FUNCTION

Doron Blatt and Alfred Hero

Department of EECS, University of Michigan, Ann Arbor, MI
{dblatt, hero}@eecs.umich.edu

ABSTRACT

Given a relative maximum of the log-likelihood function, how to assess whether it is the global maximum? In this paper we propose a statistical tool to answer this question by posing it as a hypothesis testing problem. A general framework for constructing tests for global maximum is given. The characteristics of the tests are investigated for two cases: correctly specified model and model mismatch. A finite sample approximation to the power is given, which gives a tool for performance prediction and a measure for comparison between tests. The tests are illustrated for two applications: estimating the parameters of a Gaussian mixture model and direction finding using an array of sensors - practical problems that are known to suffer from local maxima.

1. INTRODUCTION

The maximum likelihood (ML) estimation method is one of the standard tools for parameter estimation. A major drawback of this method when applied to non-linear estimation problems is the fact that the associated likelihood equations required for the derivation of the estimator rarely have a closed form analytic solution. To solve the resulting global optimization problem, initiate and converge methods are often applied. These methods are based on an initial guess (often found by a simpler method) which is followed by a local, often iterative, optimization procedure (e.g. the EM algorithm). As a consequence, the performance of these methods highly depends on the starting point. In particular, if the log-likelihood function is not strictly convex and there is no available method that is guaranteed to provide an initial guess within the attraction region of the global maximum, then there is a risk that a local search will stagnate at a local maximum. This phenomenon leads to large-scale estimation errors.

The maximum likelihood framework would benefit from an answer to the following question: Given a relative maximum of the log-likelihood function, how to assess whether this is the global maximum? In this paper we take a statistical approach to answering this question. Specifically, given a relative maximum, a statistical test is performed to test whether or not it is the global maximum.

Several global maximum tests have been proposed [1, 2, 3]. While applied to cases where the statistical model is correct, these tests are based on tests for model mismatch [4] and the observation that a local maximum of the log-likelihood function in a correctly specified model is in fact a global maximum of a misspecified model - a model in which the parameters are restricted to a region which does not contain the true parameter. A drawback of these tests is that under model mismatch, they cannot distinguish between local and global maxima.

The contribution of this paper is as follows. For correctly specified models a general framework for constructing tests that a local maximum is the global maximum is presented. Then a class of new tests is given, which are simpler to compute and in some cases give better performance than previously proposed methods. In addition, we derive an approximation of the power of the tests, which is useful for predicting performance and provides a measure for comparing between tests. For cases where model mismatch can occur, a method is given for off-line calibration of the tests to improve performance. Finally, we illustrate the method for two parameter estimation problems.

2. PROBLEM FORMULATION

Consider a collection of n i.i.d. $P \times 1$ random vectors $\mathbf{y}_t$, $t = 1, \ldots, n$ drawn from an unknown density $g(\mathbf{y})$. The information we want to extract from the data is encoded in a $K \times 1$ parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_K]^T$, through which we define a regular parametric class [5] of density functions $\{f(\mathbf{y}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

Denote by $L_n(\mathbf{Y}_n; \boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \log f(\mathbf{y}_t; \boldsymbol{\theta})$ the normalized log-likelihood function of the measurements, where $\mathbf{Y}_n = [\mathbf{y}_1\, \mathbf{y}_2\, \ldots\, \mathbf{y}_n]$. Denote by $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\mathbf{Y}_n; \boldsymbol{\theta})$ the ML estimator (MLE).

Denote by $\mathrm{E}\{\cdot\}$ the expectation with respect to true underlying density $g(\mathbf{y})$, and let $\boldsymbol{\theta}^* \stackrel{\triangle}{=} \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}\{\log f(\mathbf{y}; \boldsymbol{\theta})\}$. Theorems 2.1, 2.2, and 3.2 of White [6] assert that under possible model mismatch $\widehat{\boldsymbol{\theta}}_n \stackrel{a.s.}{\to} \boldsymbol{\theta}^*$ and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ is asymptotically zero-mean Normal distributed with covariance matrix $\mathbf{C}(\boldsymbol{\theta}^*) = \mathbf{A}^{-1}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}^{-1}(\boldsymbol{\theta}^*)$, where $\mathbf{A}(\boldsymbol{\theta}) = \mathrm{E}\left\{\nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{y}; \boldsymbol{\theta})\right\}$, $\mathbf{B}(\boldsymbol{\theta}) = \mathrm{E}\left\{\nabla_{\boldsymbol{\theta}} \log f(\mathbf{y}; \boldsymbol{\theta}) \nabla^T_{\boldsymbol{\theta}} \log f(\mathbf{y}; \boldsymbol{\theta})\right\}$. When the model is correctly specified, i.e., $g(\mathbf{y}) = f(\mathbf{y}, \boldsymbol{\theta}^0)$ for some unique $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$, this result becomes the standard consistency, and asymptotic Normality result for the MLE, where $\boldsymbol{\theta}^* = \boldsymbol{\theta}^0$, and $\mathbf{C}(\boldsymbol{\theta}^0) = -\mathbf{A}^{-1}(\boldsymbol{\theta}^0) = \mathbf{B}^{-1}(\boldsymbol{\theta}^0)$ is the inverse of the Fisher information matrix (FIM).

Denote by $\widetilde{\boldsymbol{\theta}}_n$ one of the relative maxima of the log-likelihood function. The problem addressed in this paper can be formulated as a hypothesis testing problem. Given $\widetilde{\boldsymbol{\theta}}_n$, decide between

$$\begin{aligned} \mathrm{H}_0 &: \quad \widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n \\ \mathrm{H}_1 &: \quad \widetilde{\boldsymbol{\theta}}_n \neq \widehat{\boldsymbol{\theta}}_n \ . \end{aligned} \tag{1}$$

A statistical test which gives a solution to this problem is called a *test for global maximum* [1, 2].

This research was partially supported by DARPA-MURI grant ARO DAAD 19-02-1-0262 and a Dept. of EECS Fellowship at the University of Michigan.

3. CONSTRUCTION OF THE TESTS

We start by deriving the asymptotic distribution of a general class of statistics which are functions of $\widetilde{\boldsymbol{\theta}}_n$ and $\mathbf{Y}_n$. This will lead to the construction of tests of (1). A similar treatment is given in the context of model specification testing in [4]. The tests given in [1, 2, 3] can be derived as special cases of this construction. Consider a vector valued function $\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) : \mathbb{R}^P \times \boldsymbol{\Theta} \rightarrow \mathbb{R}^Q$, and define the vectors $\mathbf{h}_n(\boldsymbol{\theta}) = 1/n \sum_{t=1}^n \mathbf{e}(\mathbf{y}_t,\boldsymbol{\theta})$, and $\mathbf{h}(\boldsymbol{\theta}) = \mathrm{E}\left\{\mathbf{e}(\mathbf{y},\boldsymbol{\theta})\right\}$, the $Q\times K$ matrix $\mathbf{H}_n(\boldsymbol{\theta}) = 1/n \sum_{t=1}^n \nabla_{\boldsymbol{\theta}}^T \mathbf{e}(\mathbf{y}_t,\boldsymbol{\theta})$, and its expectation $\mathbf{H}(\boldsymbol{\theta})$. Finally, define the $Q \times Q$ covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ by

$$\mathrm{E}\left\{\left[\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla \log f(\mathbf{y};\boldsymbol{\theta})\right] \times \left[\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla \log f(\mathbf{y};\boldsymbol{\theta})\right]^T\right\},$$

and its empirical estimate $\mathbf{V}_n(\boldsymbol{\theta})$ by

$$\frac{1}{n}\sum_{t=1}^n \left[\mathbf{e}(\mathbf{y}_t,\boldsymbol{\theta}) - \mathbf{h}_n(\boldsymbol{\theta}) - \mathbf{H}_n(\boldsymbol{\theta})\boldsymbol{A}_n^{-1}(\boldsymbol{\theta})\nabla \log f(\mathbf{y}_t;\boldsymbol{\theta})\right] \times \left[\mathbf{e}(\mathbf{y}_t,\boldsymbol{\theta}) - \mathbf{h}_n(\boldsymbol{\theta}) - \mathbf{H}_n(\boldsymbol{\theta})\boldsymbol{A}_n^{-1}(\boldsymbol{\theta})\nabla \log f(\mathbf{y}_t;\boldsymbol{\theta})\right]^T,$$

and assume that $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$ is such that $\mathbf{V}(\boldsymbol{\theta}^*)$ is nonsingular.

Theorem 1

$$n\left[\mathbf{h}_n(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta}^*)\right]^T \mathbf{V}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\left[\mathbf{h}_n(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta}^*)\right] \quad (2)$$

is asymptotically distributed as Chi-square with Q degrees of freedom (χ^2_Q).

Proofs of all theorems are given in [7]. Theorem 1 is used to construct tests for global maximum in the following manner. Choose a function $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$ having zero mean at the point $\boldsymbol{\theta}^*$, that is

$$\mathbf{h}(\boldsymbol{\theta}^*) = \mathrm{E}\left\{\mathbf{e}(\mathbf{y},\boldsymbol{\theta}^*)\right\} = \mathbf{0}_{Q\times 1} \ . \quad (3)$$

This function will be called the *global-maximum validation function*. Theorem 1 asserts that under H_0, and when (3) is satisfied, the statistic

$$S_n = n\mathbf{h}_n^T(\widetilde{\boldsymbol{\theta}}_n)\mathbf{V}_n^{-1}(\widetilde{\boldsymbol{\theta}}_n)\mathbf{h}_n(\widetilde{\boldsymbol{\theta}}_n) \ , \quad (4)$$

with $\mathbf{V}_n^{-1}(\widetilde{\boldsymbol{\theta}}_n)$ computed by (2) is asymptotically χ^2_Q distributed. Denote by $F_{\chi^2_Q}(\cdot)$ the χ^2_Q cumulative distribution function. Therefore, a false alarm level α test of the hypotheses (1) is made by comparing S_n to the threshold $F^{-1}_{\chi^2_Q}(1-\alpha)$. If S_n exceeds the threshold, H_0 is rejected and one concludes that the iterative local search should be reinitiated in the hope of convergence to a different maximum. Otherwise, the null hypothesis cannot be rejected and $\widetilde{\boldsymbol{\theta}}_n$ is declared a global maximum. If (3) does not hold for any other local maximum of the ambiguity function, then the test is consistent, i.e., it has asymptotically unit power for any $\alpha \in (0,1)$ (see [1, 3], and the discussion in Sec. 5).

3.1. Moment Based Tests

The following class of tests are based on the property that the moments of the distribution induced by the estimated parameter should be in good agreement with the empirical moments of the data. Therefore, these tests are especially suited for cases in which the underlying physical model specifies a simple parametrization of one of the moments of the data. For example, assume that the first moment of $\mathbf{y}$ is modelled by $\boldsymbol{\mu}(\boldsymbol{\theta}) = \int \mathbf{y} f(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}$, where $\boldsymbol{\mu}(\cdot)$ is a pre-specified non-linear function. Then to construct a test, which is based on the first moment, $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$ is taken to be $\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) = \mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})$. This choice of $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$ leads to $\mathbf{h}_n(\widetilde{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{t=1}^n \mathbf{y}_t - \boldsymbol{\mu}(\widetilde{\boldsymbol{\theta}}_n)$. If the first moment of the data does not depend on $\boldsymbol{\theta}$ or is weakly dependent, it is possible to base the test on any other moment. For example, one can base $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$ on one element of the correlation matrix $\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) = [\mathbf{y}]_i [\mathbf{y}]_j - \mathbf{R}_{ij}(\boldsymbol{\theta})$, where $\mathbf{R}_{ij}(\boldsymbol{\theta}) = \int [\mathbf{y}]_i [\mathbf{y}]_j f(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}$ is pre-specified from the underlying model. Tests that are based on the moments of the data are easier to compute than the tests available in the literature, and, as will be shown in the simulation results, remarkably do not reduce performance.

In Sec. 6, moment based tests are compared to Biernacki's test [3], in which $\mathbf{e}(\mathbf{y},\boldsymbol{\theta}) = \log f(\mathbf{y};\boldsymbol{\theta}) - \int \log f(\mathbf{y};\boldsymbol{\theta}) f(\mathbf{y};\boldsymbol{\theta})\mathrm{d}\mathbf{y}$. Thus, Biernacki's test compares the log-likelihood evaluated at $\widetilde{\boldsymbol{\theta}}_n$ and its expected value, which is calculated as if $\widetilde{\boldsymbol{\theta}}_n$ is the true parameter.

4. MISSPECIFIED MODELS

If the test statistic is designed under the assumption that the model is correctly specified but the actual underlying distribution is outside the parametric family, then (3) may be violated. In this case, S_n will not be χ^2_Q distributed and hence the specification of the level is incorrect. When $\mathbf{h}(\boldsymbol{\theta}^*) \neq 0$ it can be shown that the finite sample distribution of the statistic is approximately a non-central χ^2_Q with noncentrality parameter $\Delta_n = n\mathbf{h}^T(\boldsymbol{\theta}^*)\mathbf{V}^{-1}(\boldsymbol{\theta}^*)\mathbf{h}(\boldsymbol{\theta}^*)$, denoted by $\chi^2_Q(\Delta_n)$. Therefore, specifying the level of the test according to the χ^2_Q distribution is no longer valid, and in fact, as the number of samples increases, the false alarm probability increases to one regardless of the test threshold.

However, suppose an upper bound on Δ_n can be found, say μ_n. Then by setting the threshold according to the non-central Chi-square critical value $F^{-1}_{\chi^2_Q(\mu_n)}(1-\alpha)$ we insure that the false alarm probability decreases (instead of increases) with n. This will be demonstrated in Sec. 6.1.

5. FINITE SAMPLE POWER APPROXIMATION

To derive the power function, the distribution of $\widetilde{\boldsymbol{\theta}}_n$ under H_1 needs to be approximated. Therefore, assumptions on the structure of the ambiguity function, defined by $a(\boldsymbol{\theta}) = \mathrm{E}\left\{\log f(\mathbf{y};\boldsymbol{\theta})\right\}$, at different local maxima are required. Assume that the system of equations $\nabla a(\boldsymbol{\theta}) = \mathbf{0}_{K\times 1}$, has a finite number of solutions in $\boldsymbol{\Theta}$ and each one of these solutions is an interior point of $\boldsymbol{\Theta}$. In addition, at each of these points, the matrix $\nabla^2 a(\boldsymbol{\theta})$ is either negative definite or positive definite. The ambiguity function $a(\boldsymbol{\theta})$ has its global maximum at $\boldsymbol{\theta}^*$. Denote by $\boldsymbol{\theta}^m$, $m = 1,\ldots,M$, the other M local maxima of $a(\boldsymbol{\theta})$.

Theorem 2 *$\exists N$ such that $\forall n > N$, $L_n(\mathbf{Y}_n;\boldsymbol{\theta})$ has $M+1$ local maxima $w.p.1$. Furthermore, the location of these relative maxima are strongly consistent estimates for $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^m$, $m = 1,\ldots,M$.*

Let $\boldsymbol{\Theta}^m$ be a closed neighborhood of $\boldsymbol{\theta}^m$, in which $\boldsymbol{\theta}^m$ is the highest relative maximum of $a(\boldsymbol{\theta})$. Define the m'th local-MLE

by $\widehat{\boldsymbol{\theta}}_n^m = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^m} L_n(\mathbf{Y}_n;\boldsymbol{\theta})$, $m = 1,\ldots,M$. Theorem 2 asserts that for sufficiently large n, $\widetilde{\boldsymbol{\theta}}_n$ will be equal to one of the local-MLEs $\widehat{\boldsymbol{\theta}}_n^m$. The local-MLE $\widehat{\boldsymbol{\theta}}_n^m$ is the MLE associated with the model $\{f(\mathbf{y},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}^m\}$. By applying Theorem 1 we obtain the following:

Corollary 1 *If $\mathbf{V}(\boldsymbol{\theta}^m)$ is nonsingular,*

$$n\left[\mathbf{h}_n(\widehat{\boldsymbol{\theta}}_n^m) - \mathbf{h}(\boldsymbol{\theta}^m)\right]^T \mathbf{V}_n^{-1}(\widehat{\boldsymbol{\theta}}_n^m)\left[\mathbf{h}_n(\widehat{\boldsymbol{\theta}}_n^m) - \mathbf{h}(\boldsymbol{\theta}^m)\right] \quad (5)$$

is asymptotically χ^2_Q distributed.

Hence, for the test to be informative for the hypotheses, $\mathbf{h}(\boldsymbol{\theta}^m)$ must not equal $\mathbf{0}_{Q\times 1}$. Otherwise the statistic is asymptotically identically distributed under the two hypotheses. When $\widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n^m$ and $\mathbf{h}(\boldsymbol{\theta}^m) \neq \mathbf{0}_{K\times 1}$, the test statistic (4) is approximately $\chi^2_Q(\epsilon_n^m)$, where $\epsilon_n^m = n\mathbf{h}^T(\boldsymbol{\theta}^m)\mathbf{V}^{-1}(\boldsymbol{\theta}^m)\mathbf{h}(\boldsymbol{\theta}^m)$. Now, recalling that for a given level α, the threshold of the test is set to $F^{-1}_{\chi^2_Q}(1-\alpha)$, the finite sample power of the test against a local maximum at $\boldsymbol{\theta}^m$ can be approximated by

$$\beta_n \approx 1 - F_{\chi^2_Q(\epsilon_n^m)}(F^{-1}_{\chi^2_Q}(1-\alpha)) \ . \quad (6)$$

Therefore the power of the test against a local maximum at $\boldsymbol{\theta}^m$ is characterized by $\mathbf{h}^T(\boldsymbol{\theta}^m)\mathbf{V}^{-1}(\boldsymbol{\theta}^m)\mathbf{h}(\boldsymbol{\theta}^m)$, which will be called *the power characteristic of the test* as a function of $\boldsymbol{\theta}^m$. For any fixed x, $\lim_{\Delta\to\infty} F_{\chi^2_Q(\Delta)}(x) = 0$. Hence, if the power characteristic is not identically zero, the level of the test approaches 1 as n increases. In Sec. 6.1 an example will be given in which this approximation is accurate even for small n.

6. SIMULATION RESULTS

The asymptotic regime assumed throughout the paper raises the question of small sample performance. In this section, tests for global maximum are evaluated through 1000 Monte Carlo iterations. By computing the empirical level and power of the tests, we evaluate: (a) the accuracy of the asymptotic approximation $F^{-1}_{\chi^2_Q}(1-\alpha)$ for the level α threshold of the test, (b) how fast the power β_n of the test approaches 1 as the number of samples increases, and (c) how accurate is the finite sample power approximation (6). Finally, the sensitivity of the tests to model mismatch is examined and the threshold adjustment procedure of Section 4 is demonstrated.

6.1. Estimation of Gaussian Mixture Parameters

The problem of estimation of Gaussian mixture parameters arises in non-parametric density estimation [8] and a variety of clustering problems [9]. The MLE for this problem is usually found by using the EM algorithm. In [9], the authors describe a method that attempts to find the global maximum. However, even this state of the art method might stagnate at a local maximum, and therefore, tests for global maximum are useful.

We consider the univariate case, in which independent scalar measurements are generated according to a two component univariate Gaussian mixture density, where the parameter vector consists of the two means $\boldsymbol{\theta} = [\eta_1\ \eta_2]^T$. The number of components, the variances, and the mixing probabilities are assumed known. In the simulation, the true parameter is $\boldsymbol{\theta} = [0,\ 3]^T$, the variances are

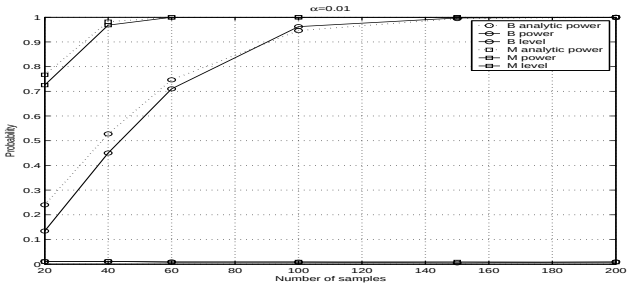

Fig. 1. Gaussian mixture: performance when the model is correctly specified.

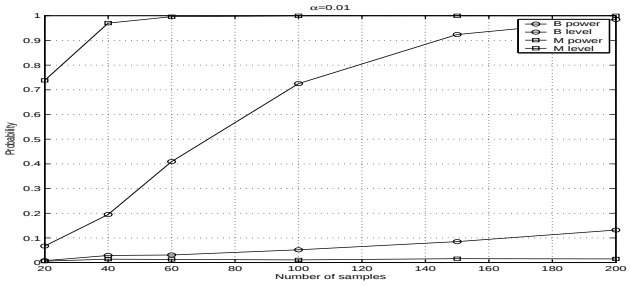

Fig. 2. Gaussian mixture: performance under model mismatch.

$\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, the mixing probabilities are $p_1 = 1 - p_2 = 0.35$ and it is known that $\boldsymbol{\Theta} = [-1,\ 4] \times [-1,\ 4]$. In this setting, the likelihood function has two relative maxima.

Biernacki's test [3] and the first moment test of Section 3.1 were applied to this problem. In Fig. 1, the empirical level and power, and the analytical approximate power (6) are presented, where B and M are shorthand notations for Biernacki's test and the first moment test, respectively.

Next, the robustness of the tests to model mismatch was evaluated. The mismatch is due to misspecified values of the parameters that are assumed known, namely the variances of the two mixtures. A discussion on scenarios in which this kind of model mismatch occurs was recently given in [10]. The MLE and the tests were computed according to the model given before but the samples were generated according to a different model. The new model, which is outside of the parametric class, is the same Gaussian mixture but with variances $\sigma_1^2 = 0.75$ and $\sigma_1^2 = 0.4$. As can be seen in Fig. 2, the moment test is robust to this model mismatch whereby Biernacki's test suffers as the number of samples increase. Biernacki's test detects the model mismatch and rejects the null hypothesis even when the relative maximum is indeed the global one. The moment test is not sensitive to this model mismatch. Even though the MLE is slightly inconsistent in this case ($\boldsymbol{\theta}^* = [-0.0248\ 3.0052]$), equation (3) is still approximately satisfied and the performance of the test is preserved.

In Fig. 3 the effect of the threshold correction of Sec. 4 is presented. An upper bound on Δ_n for each of the tests was found under the assumption that the maximal deviation from the nominal values of σ_1^2 and σ_2^2 are 0.25 and 0.1 respectively. Due to the threshold correction, the level of the tests is decreasing rather than increasing as n increases, at the price of reduced power.

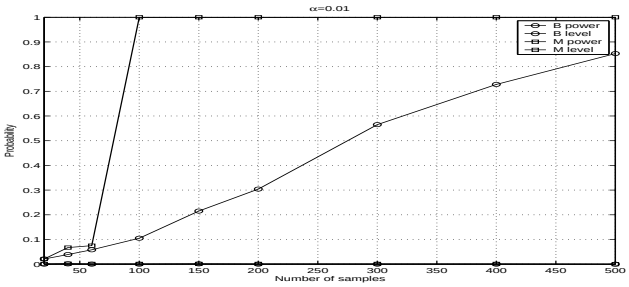

Fig. 3. Gaussian mixture: performance of the tests under model mismatch, after threshold correction.

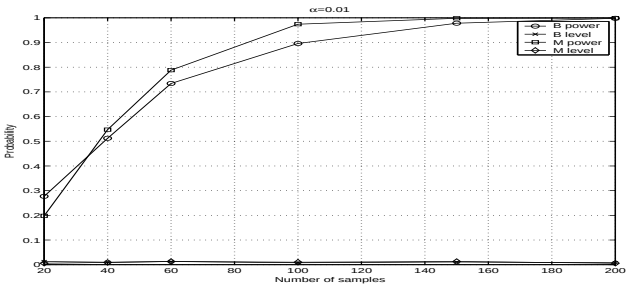

Fig. 4. Direction finding: performance when the model is correctly specified.

6.2. Direction Finding in Array Signal Processing

We adopt the standard narrow band model of [11]. We consider the estimation of the directions of two uncorrelated narrow band Gaussian sources using a uniform linear array of $P = 4$ sensors with $\lambda/2$ spacing between elements. The received signal model is given by $\mathbf{y}_t = \mathbf{D}(\boldsymbol{\theta})\mathbf{s}_t + \mathbf{e}_t$, where $\mathbf{y}_t \in \mathcal{C}^P$ is the noisy measurement vector, $\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\theta_1), \mathbf{d}(\theta_2)]$, where $[\mathbf{d}(\theta)]_p = \exp\{jp\pi\cos(\theta)\}$, $p = 0, 1, 2, 3$ is the steering vector, $\mathbf{s}_t$ contains the two signal components, and $\mathbf{e}_t$ is a temporally and spatially white circular Gaussian noise. This signal model corresponds to the so called stochastic signal model in which the received signal at the array is distributed as a temporally white zero-mean circular Gaussian random vector with covariance matrix $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{D}(\boldsymbol{\theta})\mathbf{K}_s\mathbf{D}^H(\boldsymbol{\theta}) + \sigma^2\mathbf{I}$, where, due to an uncorrelated sources assumption, $\mathbf{K}_s = diag(\sigma_{s1}^2, \sigma_{s2}^2)$, σ_{s1}^2 and σ_{s2}^2 are the two source variances, and σ^2 is the noise variance. The noise and signal variances are assumed known. The unknowns are the source directions, $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$. In the simulations $\boldsymbol{\theta} = [1.4, 1.7]^T$, $[\sigma_{s1}^2, \sigma_{s2}^2] = [1, 4]$, and $\sigma^2 = 1$. In this problem, the likelihood function has two relative maxima.

Biernacki's test [3] and a second moment test which is based on the first off diagonal element of the covariance matrix were applied. In Fig. 4 it is seen that for this choice of parameters the second moment test outperforms Biernacki's test.

Next, the robustness to model mismatch was tested as the noise variance was altered from 1 to 1.2 without changing the parametric class. In Fig. 5 it is seen that Biernacki's test is more sensitive to this kind of model mismatch than our second moment test.

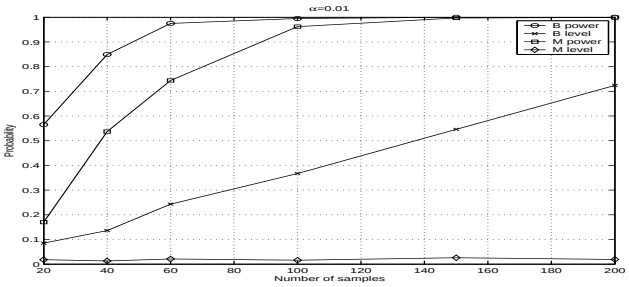

Fig. 5. Direction finding: performance under model mismatch.

7. CONCLUSIONS AND FUTURE WORK

This paper has presented a method for detecting a case in which a local search for the maximum likelihood has stagnated at a local maximum. This is a useful tool in the solution of the global optimization problem associated with the ML method. Because existing tests are sensitive to model mismatch, the general treatment given here is necessary for implementing this tool in practice. The framework given for the construction of tests and the power analysis enable us to pose fundamental questions of optimality: Given a statistical model, what is the best choice of $\mathbf{e}(\mathbf{y}, \boldsymbol{\theta})$ in terms of achieving maximum power for a given level with minimum sensitivity to model mismatch? This remains an open question.

8. REFERENCES

[1] L. Gan and J. Jiang. A test for global maximum. *Journal of the American Statistical Association*, 94:847–854, Sep 1999.

[2] R. E. Dorsey and W. J. Mayer. Detection of spurious maxima through random draw tests and specification tests. *Computational Economics*, 16:237–256, 2000.

[3] C. Biernacki. Un test pour le maximum global de vraisemblance. *35ièmes journées de statistiques, Lyon, France SFdS'2003*, June 2003.

[4] H. White. *Estimation, Inference and Specification Analysis*. Cambridge University Press, 1994.

[5] I.A. Ibragimov and R.Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, 1981.

[6] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–26, Jan. 1982.

[7] D. Blatt and A. Hero. On tests for global maximum of the likelihood function. *In preparation*.

[8] R. D. Nowak. Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE Trans. Signal Process.*, 51(8):2245–2253, August 2003.

[9] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Machine Intell.*, 24:381–396, March 2002.

[10] W. Xu, A. Baggeroer, and K. L. Bell. A bound on mean-square estimation error with background parameter mismatch. *IEEE Trans. Inform. Theory*, 50(5):621–632, 2004.

[11] P. Stoica and A. Nehorai. Music, maximum likelihood, and Cramér Rao bound. *IEEE Trans. Signal Process.*, 37(5):720–741, May 1989.