On Tests for Global Maximum of the Log-Likelihood Function

Doron Blatt, *Student Member, IEEE,* and Alfred O. Hero, III, *Fellow, IEEE*

***Abstract*— Given the location of a relative maximum of the log-likelihood function, how to assess whether it is the global maximum? This paper investigates a statistical tool, which answers 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 sensitivity of the tests to model mismatch is analyzed in terms of the Renyi divergence and the Kullback-Leibler distance between the true underlying distribution and the assumed parametric class and tests that are insensitive to small deviations from the model are derived. The tests are illustrated for three applications: passive localization or direction finding using an array of sensors, estimating the parameters of a Gaussian mixture model, and estimation of superimposed exponentials in noise - problems that are known to suffer from local maxima.**

***Index Terms*— Parameter estimation, maximum likelihood, global optimization, local maxima, array processing, Gaussian mixtures, superimposed exponentials in noise.**

I. Introduction

THE maximum likelihood (ML) estimation method is one of the standard tools for parameter estimation. Among its appealing properties are consistency and asymptotic efficiency [1]–[3]. However, 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. This shortcoming poses a global optimization problem. Solving this problem by applying numerical methods is usually computationally prohibitive. To date, there have been few global optimization methods applied to ML estimation (e.g. [4]–[8]) because of the computational complexity involved. More commonly, initiate and converge methods are 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 expectation maximization algorithm [9] and its variations [10], Fisher scoring [10], the Gauss-Newton method [11], and majorizing or minorizing algorithms [12], [13]). 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 location of a relative maximum of the log-likelihood function, how to assess whether this is the global maximum? One approach to this question is the Kronecker-Picard integral framework [6]. However, the computation of this multi-dimensional integral is difficult, indeed equivalent to the complexity involved in finding the global maximum, rendering this approach impractical. Instead, in this paper we take a statistical approach to answering this question.

The first statistical solutions for discriminating between local and global maxima were based on sampling the domain of the log-likelihood function. Given a sequence of random starting points and the corresponding set of relative maxima found by a local search method, Finch et. al. [14] proposed a statistical method to assess the probability that the global maximum has not yet been found based on an asymptotic (in the number of starting points) result on the total probability of unobserved outcomes due to Bickel and Yahav [15]. Veall [16] used an order statistic result due to de Haan [17] that characterizes the distribution of the ordered values of a smooth function, sampled at random points. Given a relative maximum, the log-likelihood function is evaluated at a large number of randomly selected points. If a point with a value larger than the value of the candidate maximum is found, then clearly it is not the global maximum. If no such point is found, de Haan's result is used to assess the probability that the relative maximum is the global one. Since these methods are based on sampling the domain of the log-likelihood function, they suffers from the curse of dimensionality and do not generalize well to high dimensional problems. Yet high dimensional problems are exactly those in which global optimization methods are computationally demanding.

Dorsey and Mayer [18] reported poor performance of Veall's method and, as an alternative, proposed to use the available methods for testing parametric models to answer the question at hand. They observed that a local maximum of the log-likelihood function is in fact a global maximum of a particular misspecified model - a model in which the parameters are restricted to a region that does not contain the true parameter. For scenarios in which the model is known to be correctly specified, these authors tested whether a relative maximum is the global one by applying a test that detects

The material in this paper will be presented in part at the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing.

This research was supported by DARPA-MURI grant ARO DAAD 19-02-1-0262, and the first author was also supported by a Department of EECS Fellowship at the University of Michigan.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (email: dblatt@umich.edu; hero@umich.edu).

model mismatch. If the result of the test leads to the conclusion that a model mismatch is likely, the hypothesis that the relative maximum is the global one is rejected. Otherwise, the relative maximum is declared the final estimate. Independently, Gan and Jiang [19] made the same observation and proposed White's information matrix test [20] as a test for global maximum. More recently, Biernacki [21], [22] proposed a new test, which is closely related to Cox's tests for separate families of hypotheses [23], [24], and showed through simulations that his new test outperforms White's information matrix test.

A drawback of the methods of [18], [19], and [22] is that they are sensitive to model mismatch. In particular, when the model is not specified correctly, the tests lose their power to distinguish between local and global maxima. In some engineering applications the statistical model is derived from the underlying physical phenomenon and deviations from this model are unlikely. In these cases, the methods can be directly applied. However, when there are uncertainties about the model, the methods [18], [19], and [22] need to be modified so as to not classify a global maximum of a misspecified model as a local maximum.

In this paper, the tests are derived under possible model mismatch. The sensitivity of the tests to model mismatch is analyzed in terms of the Renyi divergence and the Kullback-Leibler distance between the true underlying distribution and the assumed parametric class. The analysis leads to a simple threshold correction method that accounts for possible deviations from the model as long as these deviations are bounded in terms of the mentioned distances. When deviations from the model are defined in terms of an embedding in a larger parametric class, insensitivity to a Pitman drift is established by constructing tests based on a vector valued validation function that is orthogonal to the elements of the gradient of the log-likelihood function of the larger class. This construction leads to tests that are locally robust to deviations from the assumed model.

An exhaustive catalogue of all the available methods for model specification testing that might be considered as candidates for tests for global maximum is beyond the scope of this paper. Rather, this paper focuses on the class of M-tests, which includes the tests of [19] and [22] as special cases, and investigates their performance as tests for global maximum.

The problem of testing a relative maximum is related to the problem of eliminating spurious maxima in scenarios in which the ML estimator (MLE) is not necessarily consistent or may not even exist (see [25] and references therein). Although some of the results apply to that problem as well, we do not pursue this connection here.

In Sec. II, we review the properties of the MLE under a possible model mismatch and pose the problem of discriminating between local and global maxima as a statistical hypothesis testing problem. The general framework for constructing M-tests [26]–[28] is presented, and it is shown that two of the available tests in the literature are special cases of M-tests. In Sec. III, the consistency of the tests is established and an approximation of the finite sample power of the tests is derived, which is useful for predicting performance and provides a measure for comparing between tests. The problem of model mismatch is treated in Sec. IV. The effect of model mismatch is characterized in terms of the Renyi divergence and the Kullback-Leibler distance and two methods for making the tests robust to small deviations from the underlying model are given. Finally, to show the applicability of this framework, in Sec. V a Monte-Carlo evaluation of the performance of the tests is presented in terms of level and power under both correct and mismatched model.

II. Preliminaries

Let $y_t, t = 1, \ldots, n$ be a collection of n independent observations drawn from an unknown distribution G with density $g(y)$, $y \in \mathbb{R}^P$. The information we want to extract from the data is encoded in a $K \times 1$ parameter vector θ, through which we define a parametric family of densities $\{f(y,\theta) : \theta \in \Theta\}$ that are twice continuously differentiable in θ for all y. For scalar functions denote by $\nabla_\theta(\cdot)$ and $\nabla^2_\theta(\cdot)$ the column vector of partial derivatives and the Hessian matrix with respect to θ, respectively. For vector valued functions let $\nabla^T_\theta(\cdot)$ be the matrix whose (k,l) element is the partial derivative of the k'th element of the function with respect to the l'th element of θ. Assume that the elements of the matrices $\nabla_\theta \log f(y,\theta) \nabla^T_\theta \log f(y,\theta)$ and $\nabla^2_\theta \log f(y,\theta)$ are dominated by functions integrable with respect to G, for all $\theta \in \Theta$, a compact subspace of $\mathbb{R}^K$.

Denote by

$$L_n(Y_n;\theta) = \frac{1}{n}\sum_{t=1}^{n} \log f(y_t;\theta)$$

the normalized log-likelihood function of the measurements, where $Y_n = [y_1\, y_2\, \ldots\, y_n]$. The MLE[1] is defined as

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(Y_n;\theta). \tag{1}$$

Denote by $\mathrm{E}\{\cdot\}$ the expectation with respect to the true underlying distribution G, and by θ^* the minimizer of the Kullback-Leibler information, i.e.,

$$\theta^* = \arg\min_{\theta\in\Theta} \mathrm{E}\left\{\log \frac{g(y)}{f(y;\theta)}\right\} = \arg\max_{\theta\in\Theta} a(\theta)$$

where $a(\theta)$ is the ambiguity function, defined as

$$a(\theta) = \mathrm{E}\{\log f(y;\theta)\} \tag{2}$$

and assume that θ^* is a well defined unique interior point of Θ. Define the matrices

$$\begin{aligned} A(\theta) &= \mathrm{E}\left\{\nabla^2_\theta \log f(y;\theta)\right\} \\ B(\theta) &= \mathrm{E}\left\{\nabla_\theta \log f(y;\theta) \nabla^T_\theta \log f(y;\theta)\right\} \\ C(\theta) &= A^{-1}(\theta) B(\theta) A^{-1}(\theta) \end{aligned} \tag{3}$$

and assume that $A(\theta^*)$ and $B(\theta^*)$ are non-singular. Under these assumptions, Theorems 2.1, 2.2, and 3.2 of White [20] assert that

$$\widehat{\theta}_n \overset{a.s.}{\to} \theta^* \tag{4}$$

[1]Sometimes called quasi-MLE when the model is incorrect.

as $n \to \infty$, and $\widehat{\theta}_n$ is asymptotically Gaussian in the sense that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \xrightarrow{D} N\left(0, C(\theta^*)\right). \tag{5}$$

When $g(y) = f(y, \theta^0)$ almost everywhere for some unique $\theta^0 \in \Theta$, we say that the model is correctly specified and this result becomes the standard consistency, and asymptotic Normality result for the MLE. More specifically, if the elements of the matrix $\nabla_\theta^T \left[\nabla_\theta f(y,\theta) \cdot f(y,\theta)\right]$ are dominated by functions integrable with respect to ν, for all $\theta \in \Theta$, where ν is the dominating measure such that $g(y) = dG(y)/d\nu$, and the support of $f(y,\theta)$ does not depend on θ, then $C(\theta^0) = -A^{-1}(\theta^0) = B^{-1}(\theta^0)$ is the inverse of the Fisher information matrix (FIM) [3, p. 80].

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

$$\begin{aligned} \mathrm{H}_0: &\quad \widetilde{\theta}_n = \widehat{\theta}_n \\ \mathrm{H}_1: &\quad \widetilde{\theta}_n \neq \widehat{\theta}_n. \end{aligned} \tag{6}$$

A statistical test which gives a solution to this problem is called a *test for global maximum*.

A. M-Tests for Global Maximum

M-tests were proposed in an econometric context by Newey [26], Tauchen [27], and White [28] as a general way of testing the validity of parametric models (see [29, Ch. 9] as well). The tests are based on a vector valued test function

$$e(y,\theta) : \mathbb{R}^P \times \Theta \to \mathbb{R}^Q \tag{7}$$

which is chosen to satisfy

$$\int e(y,\theta) f(y,\theta) dy = 0, \quad \forall \theta \in \Theta. \tag{8}$$

Hence, given the MLE $\widehat{\theta}_n$, large values of $1/n \sum_{t=1}^n e(y_t, \widehat{\theta}_n)$ indicate that a model mismatch is likely. Small values of $1/n \sum_{t=1}^n e(y_t, \widehat{\theta}_n)$ indicate that the model is correctly specified or alternatively that the type of model mismatch is such that $g(y) \notin \{f(y,\theta) : \theta \in \Theta\}$ but

$$\int e(y,\theta^*) g(y) dy = 0. \tag{9}$$

The same framework can be used to construct tests for (6). First suppose that the model is correctly specified and that $e(y,\theta)$ is chosen to satisfy (8). Then, given a location of a relative maximum of the log-likelihood function $\widetilde{\theta}_n$, large values of $1/n \sum_{t=1}^n e(y_t, \widetilde{\theta}_n)$ indicate that it is not likely that $\widetilde{\theta}_n$ is the MLE. This directly extends to the case of model mismatch, if it is known that (9) holds.

The tests are constructed as follows. Assume that the elements of $e(y,\theta)$ are twice differentiable with respect to θ for every y, and that the elements of the vector $\nabla_\theta e(y,\theta)$ and the matrices $e(y,\theta)\nabla_\theta^T \log f(y,\theta)$ and $e(y,\theta)e^T(y,\theta)$ are dominated by functions integrable with respect to G for all $\theta \in \Theta$. Define the vectors

$$\begin{aligned} h_n(\theta) &= \frac{1}{n}\sum_{t=1}^n e(y_t,\theta) \\ h(\theta) &= \mathrm{E}\left\{e(y,\theta)\right\} \end{aligned} \tag{10}$$

and the $Q \times K$ matrices

$$\begin{aligned} H_n(\theta) &= \frac{1}{n}\sum_{t=1}^n \nabla_\theta^T e(y_t,\theta) \\ H(\theta) &= \mathrm{E}\left\{\nabla_\theta^T e(y,\theta)\right\}. \end{aligned} \tag{11}$$

Define the $Q \times Q$ matrix $V(\theta)$ by

$$\begin{aligned} \mathrm{E}\Big\{\big[e(y,\theta) - h(\theta) - H(\theta)A^{-1}(\theta)\nabla_\theta \log f(y;\theta)\big] \times \\ \big[e(y,\theta) - h(\theta) - H(\theta)A^{-1}(\theta)\nabla_\theta \log f(y;\theta)\big]^T\Big\} \end{aligned} \tag{12}$$

and its empirical estimate by

$$\begin{aligned} V_n(\theta) = \frac{1}{n}\sum_{t=1}^n \\ \big[e(y_t,\theta) - h_n(\theta) - H_n(\theta)A_n^{-1}(\theta)\nabla_\theta \log f(y_t;\theta)\big] \times \\ \big[e(y_t,\theta) - h_n(\theta) - H_n(\theta)A_n^{-1}(\theta)\nabla_\theta \log f(y_t;\theta)\big]^T \end{aligned} \tag{13}$$

where

$$A_n(\theta) = \frac{1}{n}\sum_{t=1}^n \nabla_\theta^2 \log f(y_t;\theta) \tag{14}$$

and assume that $e(y,\theta)$ is such that $V(\theta^*)$ in (12) is nonsingular. Under the assumptions made above,

$$\sqrt{n}\left[h_n(\widehat{\theta}_n) - h(\theta^*)\right] \xrightarrow{D} N\left(0, V(\theta^*)\right) \tag{15}$$

$$V_n(\widehat{\theta}_n) \xrightarrow{a.s.} V(\theta^*) \tag{16}$$

element by element, $V_n(\widehat{\theta}_n)$ is nonsingular for sufficiently large n, and as a result,

$$n\left[h_n(\widehat{\theta}_n) - h(\theta^*)\right]^T V_n^{-1}(\widehat{\theta}_n)\left[h_n(\widehat{\theta}_n) - h(\theta^*)\right] \tag{17}$$

is asymptotically Chi-Squared distributed with Q degrees of freedom [26]–[28]. An elementary proof of this result is included in the Appendix for completeness.

Based on this result, tests for global maximum can be constructed as follows. Choose a function $e(y,\theta)$ having mean zero at the point θ^*, that is

$$h(\theta^*) = \mathrm{E}\left\{e(y,\theta^*)\right\} = 0. \tag{18}$$

The function $e(y,\theta)$ will be called the *global maximum validation function*. Under H_0 and when (18) is satisfied, the statistic

$$S_n = n h_n^T(\widetilde{\theta}_n) V_n^{-1}(\widetilde{\theta}_n) h_n(\widetilde{\theta}_n) \tag{19}$$

with $V_n^{-1}(\widetilde{\theta}_n)$ computed by (13) is asymptotically Chi-Squared distributed with Q degrees of freedom, denoted by χ_Q^2. Denote by $F_{\chi_Q^2}(\cdot)$ the χ_Q^2 cumulative distribution function. Therefore, a false alarm level α test of the hypotheses (6) is made by comparing S_n to $F_{\chi_Q^2}^{-1}(1-\alpha)$, which is the critical value of the χ_Q^2 distribution for the desired false alarm level. If

S_n exceeds the critical value, H_0 is rejected and one concludes that the iterative local search should be re-initiated in the hope of convergence to a different maximum. Otherwise, the null hypothesis cannot be rejected and $\widetilde{\theta}_n$ is declared the final estimate.

When the model is correctly specified, $\theta^* = \theta^0$ and Eq. (18) becomes

$$h(\theta^0) = \mathrm{E}\left\{e(y,\theta^0)\right\} = \int e(y,\theta^0) f(y,\theta^0) dy = 0. \quad (20)$$

A global maximum validation function $e(y,\theta)$ satisfying (20) can be constructed from any random function, e.g. call it $\overline{e}(y,\theta)$, by replacing it with the centered statistic:

$$e(y,\theta) = \overline{e}(y,\theta) - \int \overline{e}(y,\theta) f(y;\theta) dy. \quad (21)$$

This construction ensures that the mean of the validation function at the true parameter is zero. Under this construction, $h_n(\widetilde{\theta}_n)$ (10) becomes

$$h_n(\widetilde{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} \overline{e}(y_t,\widetilde{\theta}_n) - \int \overline{e}(y,\widetilde{\theta}_n) f(y;\widetilde{\theta}_n) dy \quad (22)$$

and the property $h(\theta^0) = \mathrm{E}\left\{e(y,\theta^0)\right\} = 0$ holds. This manipulation requires an analytical solution of the integral in (22) or its approximation via numerical integration.

Two tests for global maximum that are available in the literature fall into this framework. Taking $e(y,\theta)$ to be the vector valued function defined as

$$[e(y,\theta)]_q = \frac{\partial^2 \log f(y;\theta)}{\partial\theta_{i_q}\partial\theta_{j_q}} + \frac{\partial \log f(y;\theta)}{\partial\theta_{i_q}} \frac{\partial \log f(y;\theta)}{\partial\theta_{j_q}} \quad (23)$$

where $[\cdot]_q$ denotes the vector's q'th element, and the indices i_q and j_q, $q = 1,\ldots,Q$, are chosen so that $V(\theta^*)$ is nonsingular, we obtain White's information matrix test [20] which was used by Gan and Jiang as their test for global maximum [19]. This test is motivated by the fact that when the model is correctly specified, $A_n(\widehat{\theta}_n)$ defined in (14), and $B_n(\widehat{\theta}_n)$, defined by

$$B_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \nabla_\theta \log f(y_t;\theta) \nabla_\theta^T \log f(y_t;\theta) \quad (24)$$

converge a.s. as $n \to \infty$ to the -FIM and FIM, respectively; an idea that was originally used by White in his test for model mismatch [20]. Hence, when the model is correctly specified, (18) is satisfied since the expected value of the sum at θ^0 vanishes. Gan and Jiang noted that White's test suffers from slow convergence rates to unit power, i.e., it requires a large number of samples to detect local maxima with high probability. A test with better convergence rates was recently proposed by Biernacki [22]. The cost of this improvement is increased complexity due to the need to evaluate an integral of the type (22). The validation function $e(y,\theta)$ associated with Biernacki's test is the scalar function

$$e(y,\theta) = \log f(y;\theta) - \int \log f(y;\theta) f(y;\theta) dy \quad (25)$$

which is a special case of (21). Hence,

$$h_n(\widetilde{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} \log f(y_t;\widetilde{\theta}_n) - \int \log f(y;\widetilde{\theta}_n) f(y;\widetilde{\theta}_n) dy. \quad (26)$$

This test is closely related to Cox's tests of separate families of hypotheses [23], [24]. The choice (25) of $e(y,\theta)$ leads to a test that compares the log-likelihood evaluated at $\widetilde{\theta}_n$ to its expected value, which is calculated as if $\widetilde{\theta}_n$ is the true parameter. The test requires the evaluation of an integral (26) of dimension P - the dimension of y. This might be prohibitive in real time applications, although in Sec. V-A below, a closed form expression for the case of Gaussian distributed y_t is given. In [21], [22] the variance estimator required for the construction of S_n (19) is consistent for $\mathrm{E}\left\{e(y,\theta^0)e^T(y,\theta^0)\right\}$ rather than for $V(\theta^0)$ (12). From (28) below, it can be seen that under the null hypothesis H_0 and when the model is correctly specified, $\mathrm{E}\left\{e(y,\theta^0)e^T(y,\theta^0)\right\}$ is an upper bound on the asymptotic variance of $\sqrt{n}h_n(\widetilde{\theta}_n)$ (26). The bound is tight when either $B(\theta^0)$ is large, e.g., at high signal to noise ratio, or when $H(\theta^0)$ is small, i.e., the expectation of the gradient of $e(y,\theta)$ is small, but in general the variance estimator of [21], [22] leads to a test with a false alarm level smaller than the specified value.

B. Moments Matching Tests

Moments matching tests were previously proposed as tests for model mismatch (see e.g. [27]) but were not applied to the problem of discrimination of local maxima. The 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 mean of y is modelled by $\mu(\theta)$, i.e. $\mu(\theta) = \int y f(y;\theta) dy$, where $\mu(\cdot)$ is a pre-specified non-linear function, then to construct a test, which is based on the first moment, $e(y,\theta)$ is taken to be

$$e(y,\theta) = y - \mu(\theta).$$

This choice of $e(y,\theta)$ leads to the empirical estimate

$$h_n(\widetilde{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} y_t - \mu(\widetilde{\theta}_n).$$

It is clear that under a correctly specified model, equation (18) is satisfied. If the model is not correctly specified but the specification of the mean is correct, the condition

$$h(\theta^*) = \mathrm{E}\{y\} - \mu(\theta^*) = 0 \quad (27)$$

will still hold if the parametric class $\{f(y;\theta) : \theta \in \Theta\}$ belongs to the linear exponential family [29].

If the mean of the data does not depend on θ or is weakly dependent, one can improve the test by including higher order moments. For example, one can specify $e(y,\theta)$ as one or more elements of the difference between sample and ensemble covariance matrices:

$$[e(y,\theta)]_q = [y]_{i_q}[y]_{j_q} - [R(\theta)]_{i_q,j_q}, \quad q = 1,\ldots,Q$$

where for matrices $[\cdot]_{q,k}$ denotes the (q,k) element, and $[R(\theta)]_{i_q,j_q} = \int [y]_{i_q}[y]_{j_q} f(y;\theta) dy$ is pre-specified from the underlying model.

C. Covariance Matrix Estimation

It is possible to exploit properties of the null hypothesis H_0 (6) in order simplify and improve the estimator (13) of the covariance matrix of $\sqrt{n}h_n(\widetilde{\theta}_n)$ (see e.g. [19], [20], [26], [27], [29]). Under H_0 $\sqrt{n}h_n(\widetilde{\theta}_n)$ equals $\sqrt{n}h_n(\widehat{\theta}_n)$, and since by construction $h(\theta^*)=0$, it is possible to drop the term $h_n(\widetilde{\theta}_n)$, which appears in (13) after substituting $\widetilde{\theta}_n$. Furthermore, when the model is correctly specified, under H_0, the asymptotic covariance matrix of $\sqrt{n}h_n(\widetilde{\theta}_n)$ simplifies to

$$\mathrm{E}\left\{e(y,\theta^0)e^T(y,\theta^0)\right\} - H(\theta^0)B^{-1}(\theta^0)H^T(\theta^0) \qquad (28)$$

where $B(\theta)$ and $H(\theta)$ are given in (3) and (11), respectively, and since a correct model specification is assumed, expectations are taken with respect to the density $f(y,\theta^0)$. Using this property, the following covariance estimators can be considered. The first is based on the data and the form (28):

$$\begin{aligned}\widehat{V}_n(\widetilde{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} e(y_t,\widetilde{\theta}_n)e^T(y_t,\widetilde{\theta}_n) \qquad (29)\\ - H_n(\widetilde{\theta}_n)B_n^{-1}(\widetilde{\theta}_n)H_n^T(\widetilde{\theta}_n)\end{aligned}$$

where $B_n(\theta)$ and $H_n(\theta)$ are defined in (24) and (11), respectively. In the correct model case, under H_0 the estimator (29) converges a.s. to the covariance matrix (28) [30, Lemma 3.1], and hence it is positive definite a.s. for sufficiently large n. The second estimator is given by

$$\begin{aligned}\overline{V}_n(\widetilde{\theta}_n) = \int e(y,\widetilde{\theta}_n)e^T(y,\widetilde{\theta}_n)f(y,\widetilde{\theta}_n)dy - \qquad (30)\\ \overline{H}(\widetilde{\theta}_n)\overline{B}^{-1}(\widetilde{\theta}_n)\overline{H}^T(\widetilde{\theta}_n)\end{aligned}$$

where

$$\overline{B}(\theta) = \int \nabla_\theta \log f(y;\theta)\nabla_\theta^T \log f(y;\theta) f(y;\theta)dy$$

and

$$\overline{H}(\theta) = \int \nabla_\theta^T e(y,\theta) f(y;\theta)dy.$$

It should be noted that under H_1 or under model mismatch, these estimates are not necessarily consistent and the estimator (29) is not necessarily positive definite.

A number of authors investigated ways of estimating the covariance matrix in scenarios in which unexpected dependencies between the measurements may occur (see e.g. [29], [31] and references therein). Methods for eliminating the requirement for covariance matrix estimation altogether were recently proposed in [32] for the problem of model testing in non-linear regression.

III. Power Analysis

In order to derive the power function, the asymptotic distribution of $\widetilde{\theta}_n$ under H_1 needs to be determined. Therefore, assumptions on the structure of the ambiguity function (2) at different local maxima are required. Assume that the system of equations $\nabla a(\theta)=0$, has a finite number of solutions in Θ and each one of these solutions is an interior point of Θ. In addition, at each of these points, the matrix $\nabla^2 a(\theta)$ is either negative definite or positive definite. The ambiguity function $a(\theta)$ has its global maximum at θ^*; denote by θ^m, $m=1,\ldots,M$, the other M local maxima of $a(\theta)$.

Theorem 1: For sufficiently large n, $L_n(Y_n;\theta)$ has $M+1$ local maxima for almost every sequence $\{y_t\}_{t\geq 1}$. Furthermore, the location of these relative maxima are strongly consistent estimates for θ^* and θ^m, $m=1,\ldots,M$.

Proof: The outline of the proof goes as follows. First we prove that, for sufficiently large n, the norm of the first derivative vector of $L_n(Y_n;\theta)$ is strictly positive outside of arbitrary small neighborhoods of the local maxima and local minima of $a(\theta)$. Then, we prove that when restricted to these neighborhoods, $L_n(Y_n;\theta)$ is either strictly convex or strictly concave and hence has a single minimum or a single maximum, respectively.

Under the assumptions made, [33, Thm. 2] gives the following uniform strong law of large numbers:

$$\begin{aligned}L_n(Y_n;\theta) &\to \mathrm{E}\left\{\log f(y;\theta)\right\} \qquad (31)\\ \nabla_\theta L_n(Y_n;\theta) &\to \mathrm{E}\left\{\nabla_\theta \log f(y;\theta)\right\}\\ \nabla_\theta^2 L_n(Y_n;\theta) &\to \mathrm{E}\left\{\nabla_\theta^2 \log f(y;\theta)\right\}\end{aligned}$$

as $n\to\infty$ uniformly in Θ for almost every sequence $\{y_t\}_{t\geq 1}$.

Denote the relative minimum points for the ambiguity function by $\phi^j\in\Theta$, $j=1,\ldots,J$, $J\geq 0$. By the assumption, $\nabla_\theta a(\theta)=0$ at the points θ^*, $\theta^m, m=1,\ldots,M$ and $\phi^j, j=1,\ldots,J$ and only at these points. In addition, the matrix $\nabla^2 a(\theta)$ is negative definite at the points θ^*, $\theta^m, m=1,\ldots,M$ and positive definite at the points $\phi^j, j=1,\ldots,J$. Denote the eigenvalues of the matrix $\nabla^2 a(\theta)$ by $\lambda_k(\theta), k=1,\ldots,K$. Therefore,

$$\begin{aligned}&\max_k\{\lambda_k(\theta^*)\}<0\\ &\max_k\{\lambda_k(\theta^m)\}<0, \quad \forall m=1,\ldots,M\end{aligned}$$

and

$$\min_k\{\lambda_k(\phi^j)\}>0, \quad \forall j=1,\ldots,J.$$

The eigenvalues are continuous functions of the matrix element and the operations $\max$ and $\min$ are also continuous in their arguments. Therefore, there are disjoint open neighborhoods $\mathcal{N}^*$, $\mathcal{N}^m$, and $\mathcal{M}^j$ around θ^*, θ^m and ϕ^j, respectively, $m=1,\ldots,M$, $j=1,\ldots J$, that satisfy the following conditions:

$$\begin{aligned}&\sup_{\theta\in\mathcal{N}^*}\max_k\{\lambda_k(\theta)\}\leq\overline{\delta}<0 \qquad (32)\\ &\sup_{\theta\in\mathcal{N}^m}\max_k\{\lambda_k(\theta)\}\leq\overline{\delta}<0, \quad \forall m=1,\ldots,M\\ &\inf_{\theta\in\mathcal{M}^j}\min_k\{\lambda_k(\theta)\}\geq\underline{\delta}>0, \quad \forall j=1,\ldots,J.\end{aligned}$$

Denote

$$\widetilde{\Theta} = \Theta \setminus \left[\mathcal{N}^* \bigcup \left(\bigcup_{m=1}^{M}\mathcal{N}^m\right) \bigcup \left(\bigcup_{j=1}^{J}\mathcal{M}^j\right)\right].$$

Since $\widetilde{\Theta}$ is also compact, and $|\partial a(\theta)/\partial\theta_k|$ is bounded and continuous for all k, we have

$$\inf_{\theta\in\widetilde{\Theta}}\sum_{k=1}^{K}|\partial a(\theta)/\partial\theta_k| = \min_{\theta\in\widetilde{\Theta}}\sum_{k=1}^{K}|\partial a(\theta)/\partial\theta_k| = \delta.$$

Since by the assumption all the stationary points of $a(\theta)$ are outside of $\widetilde{\Theta}$, δ is strictly positive.

Next, we prove that there exist N_1 such that $\forall n > N_1$,

$$\sum_{k=1}^{K} |\partial L_n(Y_n;\theta)/\partial\theta_k| > \delta/2, \quad \forall\theta \in \widetilde{\Theta}, \quad w.p.\ 1$$

i.e., for sufficiently large n, the function $L_n(Y_n;\theta)$ has no stationary points in $\widetilde{\Theta}$ for almost every sequence $\{y_t\}_{t\geq 1}$. To this end, choose N_1 such that for all $n > N_1$,

$$|\partial a(\theta)/\partial\theta_k - \partial L_n(Y_n;\theta)/\partial\theta_k| < \frac{\delta}{2K},$$
$$\forall k = 1,\ldots,K,\ \forall\theta \in \widetilde{\Theta}, \quad w.p.\ 1$$

which can always be found by (31). Therefore,

$$\sum_{k=1}^{K} |\partial a(\theta)/\partial\theta_k - \partial L_n(Y_n;\theta)/\partial\theta_k| < \frac{\delta}{2},$$
$$\forall\theta \in \widetilde{\Theta}, \quad w.p.\ 1$$

and hence, $\forall n > N_1$,

$$\sum_{k=1}^{K} |\partial L_n(Y_n;\theta)/\partial\theta_k| > \frac{\delta}{2}, \quad \forall\theta \in \widetilde{\Theta}, \quad w.p.\ 1$$

and the claim is proved.

Next, we prove that there exist N_2 such that $\forall n > N_2$, $L_n(Y_n;\theta)$ is concave over $\overline{\mathcal{N}}^*$, $\overline{\mathcal{N}}^m, m = 1,\ldots,M$ and convex over $\overline{\mathcal{M}}^j, j = 1,\ldots,J$, where $\overline{\mathcal{N}}$ denotes the closure of the set $\mathcal{N}$. Denote the eigenvalues of $\nabla^2 L_n(Y_n;\theta)$ by $\lambda_k^n(\theta), k = 1,\ldots,L$. We consider one specific neighborhood $\overline{\mathcal{N}}^*$, and prove that

$$\max_{\theta\in\overline{\mathcal{N}}^*} \max_k \{\lambda_k^n(\theta)\} < \frac{\overline{\delta}}{2} < 0, \quad \forall n > N_2, \quad w.p.\ 1 \qquad (33)$$

where $\overline{\delta}$ was defined in (32), i.e., $L_n(Y_n;\theta)$ is concave over $\overline{\mathcal{N}}^*$.

By the construction, the maximal eigenvalue is uniformly continuous over $\overline{\mathcal{N}}^*$. Therefore,

$$\max_k \{\lambda_k^n(\theta)\} \to \max_k \{\lambda_k(\theta)\}, \quad \forall\theta \in \overline{\mathcal{N}}^*, \quad w.p.\ 1$$

and (33) follows. The same argument holds for the proof of concavity of $L_n(Y_n;\theta)$ over the rest of the neighborhoods $\overline{\mathcal{N}}^m$, $m = 1,2,\ldots,M$ and the convexity of $L_n(Y_n;\theta)$ over $\overline{\mathcal{M}}^j$, $j = 1,\ldots,J$.

For each set $\overline{\mathcal{N}}^m$, by (31) as n increases $L_n(Y_n;\theta)$ will eventually be greater at θ^m than at any point on the boundary of $\overline{\mathcal{N}}^m$, $w.p.\ 1$. Therefore, $L_n(Y_n;\theta)$ will attain a single local maximum at an interior point of $\mathcal{N}^m$, $w.p.\ 1$ (not necessarily at θ^m). A similar argument holds for $\overline{\mathcal{N}}^*$ and for a minimum point in $\mathcal{M}^j$ and the first part of the theorem is proved.

Finally, since the sets $\mathcal{N}^*$, $\mathcal{N}^m, m = 1,\ldots,M$ can be taken arbitrarily small, the maximum points of $L_n(Y_n;\theta)$ are strongly consistent estimates of θ^*, $\theta^m, m = 1,\ldots,M$. ■

Theorem 1 ensures that as n increases the relative maxima of the log-likelihood function occur close to the relative maxima of the ambiguity function and only at these locations. This implies that the relative maxima of the log-likelihood function are asymptotically Gaussian distributed. More specifically, let Θ^m be a closed neighborhood of θ^m, in which θ^m is the highest relative maximum of $a(\theta)$. Define the m'th local-MLE by

$$\widehat{\theta}_n^m = \arg\max_{\theta\in\Theta^m} L_n(Y_n;\theta), \quad m = 1,\ldots,M. \qquad (34)$$

If the optimization method used to solve (1) is certain to find a relative maximum of $L_n(Y_n;\theta)$, then Theorem 1 asserts that for sufficiently large n, $\widetilde{\theta}_n$ will be equal to one of the local-MLEs $\widehat{\theta}_n^m$, w.p. 1. The local-MLE $\widehat{\theta}_n^m$ is the MLE associated with the model $\{f(y,\theta) : \theta \in \Theta^m\}$ and therefore falls into the mismatch model framework of White [20]. Hence we have the following.

Corollary 1: For all m:

1) $\widehat{\theta}_n^m \overset{a.s.}{\to} \theta^m$ as $n \to \infty$, and
2) $\sqrt{n}\left(\widehat{\theta}_n^m - \theta^m\right) \overset{D}{\to} N\left(0, C(\theta^m)\right)$.

In addition, by (15)-(17) we obtain the following:

Corollary 2: For all m:

$$\sqrt{n}\left[h_n(\widehat{\theta}_n^m) - h(\theta^m)\right] \overset{D}{\to} N\left(0, V(\theta^m)\right)$$

$V_n(\widehat{\theta}_n^m) \overset{a.s.}{\to} V(\theta^m)$ element by element. In addition, assuming that $V(\theta^m)$ is nonsingular,

$$n\left[h_n(\widehat{\theta}_n^m) - h(\theta^m)\right]^T V_n^{-1}(\widehat{\theta}_n^m)\left[h_n(\widehat{\theta}_n^m) - h(\theta^m)\right] \qquad (35)$$

is asymptotically distributed as χ_Q^2.

From Corollary 2 it is clear that for the test to have power against $\widehat{\theta}_n^m$, $h(\theta^m)$ must not equal 0. Otherwise the statistic has the same asymptotic χ_Q^2 distribution under both hypotheses H_0 and H_1 (6). On the other hand, if $h(\theta^m) \neq 0$ the consistency of the test can be established.

Corollary 3: Assume $\widetilde{\theta}_n = \widehat{\theta}_n^m$. If $h(\theta^m) \neq 0$ then

$$\Pr\{S_n > F_{\chi_Q^2}^{-1}(1-\alpha)\} \to 1$$

for every choice of level $\alpha \in (0,1)$.

Proof: Under the assumption, $h_n(\widetilde{\theta}_n) \overset{a.s.}{\to} h(\theta^m)$ by [**?**, Lemma 3.1]. Therefore, since $V_n(\widehat{\theta}_n^m) \overset{a.s.}{\to} V(\theta^m)$ element by element and we assumed that $V(\theta^m)$ is nonsingular,

$$\Pr\{S_n > \varepsilon\} \to 1$$

for all $\varepsilon > 0$, by [29, Thm. 8.13]. ■

Implied from corollary 3 is the consistency of the test: If $h(\theta^m) \neq 0$ for all $m = 1,\ldots,M$, then

$$\Pr\{S_n > F_{\chi_Q^2}^{-1}(1-\alpha)|\mathrm{H}_1\} \to 1 \qquad (36)$$

for every choice of level $\alpha \in (0,1)$, i.e., the test is consistent. This result extends the results of [19] and [22], which established under a correctly specified model (each for their own global maximum validation function) that if the only solution to the set of equations

$$\int \nabla_\theta \log f(y,\theta) f(y,\theta^0) dy = 0$$
$$\int e(y,\theta) f(y,\theta^0) dy = 0$$

is θ^0, then

$$\sqrt{n} h_n(\widetilde{\theta}_n) \overset{D}{\to} N(0, V(\theta^0)) \quad \text{iff} \quad \widetilde{\theta}_n = \widehat{\theta}_n.$$

Furthermore, Corollary 2 implies that under H_1, and particularly when $\widetilde{\theta}_n = \widehat{\theta}_n^m$, the distribution of the test statistic S_n is approximately non-central χ^2_Q with non-centrality parameter

$$n\delta^m = nh^T(\theta^m)V^{-1}(\theta^m)h(\theta^m)$$

denoted by $\chi^2_Q(n\delta^m)$ [34]. We denote the $\chi^2_Q(n\delta^m)$ cumulative distribution function by $F_{\chi^2_Q(n\delta^m)}(\cdot)$. The finite sample power of the test against a local maximum at θ^m can be approximated by [34, p. 468]

$$1 - F_{\chi^2_Q(n\delta^m)}\left[F^{-1}_{\chi^2_Q}(1-\alpha)\right]. \tag{37}$$

Therefore, the power of a given test against a local maximum at θ^m is characterized by

$$\delta^m = h^T(\theta^m)V^{-1}(\theta^m)h(\theta^m) \tag{38}$$

which will be called the power characteristic of the test as a function of θ^m. The power characteristic is a basis of comparison between tests.

IV. Misspecified Models

In general, it is difficult to discriminate between the cases of: (a) $\widetilde{\theta}_n$ a local maximum in a correctly specified model; and (b) $\widetilde{\theta}_n$ a global maximum in a misspecified model. Under model mismatch, the probability of mistakenly rejecting $\widetilde{\theta}_n$ as the global maximum, increases with the number of samples.

If the test statistic is designed under the assumption that the model is correctly specified but the actual underlying distribution is outside the assumed parametric family, then (18) may be violated. In this case, even when $\widetilde{\theta}_n = \widehat{\theta}_n$, $h_n(\widetilde{\theta}_n) \overset{a.s.}{\to} h(\theta^*) \neq 0$ and, similar to the discussion in the previous section, S_n is approximately distributed as $\chi^2_Q(n\epsilon)$ with non-centrality parameter $n\epsilon = nh^T(\theta^*)V^{-1}(\theta^*)h(\theta^*)$, instead of the assumed central chi-squared. In this case, as n tends to infinity, the probability of mistakenly rejecting $\widetilde{\theta}_n$ as the global maximum increases to one regardless of the test threshold, and is approximately given by

$$1 - F_{\chi^2_Q(n\epsilon)}\left[F^{-1}_{\chi^2_Q}(1-\alpha)\right].$$

A. A Bound on the Non-Centrality Parameter

It is possible to bound the non-centrality parameter ϵ, induced by the model mismatch, in terms of the Renyi divergence between $f(y;\theta^*)$ and true underlying density $g(y)$. Consider the case in which $e(y,\theta)$ is a scalar function and satisfies

$$\int e(y,\theta)f(y,\theta)dy = 0, \quad \forall\theta \in \Theta.$$

In this case the non-centrality parameter simplifies to

$$n\epsilon = nh^2(\theta^*)/V(\theta^*).$$

Since θ^* minimizes $D\left(g(y)||f(y,\theta)\right)$ with respect to θ,

$$\int \nabla_\theta \log f(y,\theta)|_{\theta=\theta^*}\, g(y)dy = 0.$$

Therefore, denoting

$$d(y,\theta) = e(y,\theta) - H(\theta)A^{-1}(\theta)\nabla^T_\theta \log f(y,\theta)$$

we obtain

$$\begin{aligned} h(\theta^*) &= \mathrm{E}\left\{e(y,\theta^*)\right\} = \mathrm{E}\left\{d(y,\theta^*)\right\} \\ &= \int \left[d(y,\theta^*) - h(\theta^*)\right]\left[g(y) - f(y,\theta^*)\right]dy. \end{aligned}$$

By the Cauchy-Schwartz inequality

$$\begin{aligned} h^2(\theta^*) \leq &\int \left[d(y,\theta^*) - h(\theta^*)\right]^2 g(y)dy \times \\ &\int \frac{\left[g(y) - f(y,\theta^*)\right]^2}{g(y)}dy \\ &= V(\theta^*)\left(\int \frac{f^2(y,\theta^*)}{g(y)}dy - 1\right) \end{aligned}$$

implying that

$$\epsilon = \frac{h^2(\theta^*)}{V(\theta^*)} \leq \exp\left[D_2\left(f(y,\theta^*)||g(y)\right)\right] - 1$$

where

$$D_\alpha(f_1(y)||f_2(y)) = \frac{1}{\alpha-1}\log\int f_1^\alpha(y)f_2^{1-\alpha}(y)dy$$

is the Renyi divergence between $f_1(y)$ and $f_2(y)$ with parameter α.

Therefore, when a bound on $D_2\left(f(y,\theta^*)||g(y)\right)$ is available, say B_ϵ, it is possible to set the threshold of the test according to a $\chi^2_Q(n\left[\exp(B_\epsilon) - 1\right])$ distribution, i.e., reject the null hypothesis if

$$S_n > F^{-1}_{\chi^2_Q(n[\exp(B_\epsilon)-1])}(1-\alpha). \tag{39}$$

This choice of threshold leads to a test, the level of which decreases to zero, instead of increasing to one. Since

$$F^{-1}_{\chi^2_Q(n[\exp(B_\epsilon)-1])}(1-\alpha) > F^{-1}_{\chi^2_Q}(1-\alpha)$$

for all α [34], this adjustment decreases the power of the test. However, as long as the the power characteristic of the test at a local maximum δ^m (38) is larger than $\exp(B_\epsilon) - 1$, the test will detect such a local maximum with probability approaching one as n tends to infinity.

Often it is difficult to compute a bound on $D_2\left(f(y,\theta^*)||g(y)\right)$, especially due to the computation required for θ^*. When the true underlying distribution and the assumed parametric model are both embedded in a larger parametric class and are sufficiently close to one another, it is possible to approximate the Renyi divergence by the Kullback-Leibler distance defined below. This leads to a simple approximation of B_ϵ.

Suppose that the parametric class $\{f(y;\theta) : \theta \in \Theta\}$ is embedded in a larger class $\{\widetilde{f}(y;\theta,\gamma) : \theta \in \Theta,\ \gamma \in \Gamma \subset \mathbb{R}^{K'}\}$ such that $f(y;\theta) = \widetilde{f}(y;\theta,\gamma^0)$ for all $\theta \in \Theta$, and that the true underlying density is $g(y) = \widetilde{f}(y;\theta^0,\gamma^1)$, with θ^1 close to θ^0. This setting was recently treated in [35], where the parameter vector γ was referred to as the background parameter.

In this case, the local equivalence and symmetry of f-divergence measures [36, p. 85] can be used to approximate the Renyi divergence

$$D_2\left(f(y,\theta^*)||g(y)\right) = D_2\left(\widetilde{f}(y;\theta^*,\gamma^0)||\widetilde{f}(y;\theta^0,\gamma^1)\right)$$

by

$$2D_1\left(\widetilde{f}(y;\theta^0,\gamma^1)||\widetilde{f}(y;\theta^*,\gamma^0)\right)$$

up to terms of order $O\left(||\theta^*-\theta^0||^3+||\gamma^0-\gamma^1||^3\right)$, where

$$\begin{aligned}D_1\left(f_1(y)||f_2(y)\right) &= \lim_{\alpha\to 1} D_\alpha\left(f_1(y)||f_2(y)\right)\\ &= \int \log\left(\frac{f_1(y)}{f_2(y)}\right) f_1(y)dy\end{aligned}$$

is the Kullback-Leibler distance between $f_1(y)$ and $f_2(y)$.

Furthermore, θ^* minimizes $D_1\left(\widetilde{f}(y;\theta^0,\gamma^1)||\widetilde{f}(y;\theta,\gamma^0)\right)$ over $\theta\in\Theta$. Hence,

$$\begin{aligned}D_1\left(\widetilde{f}(y;\theta^0,\gamma^1)||\widetilde{f}(y;\theta^*,\gamma^0)\right) \leq\\ D_1\left(\widetilde{f}(y;\theta^0,\gamma^1)||\widetilde{f}(y;\theta^0,\gamma^0)\right).\end{aligned}$$

Therefore, $D_2\left(f(y,\theta^*)||g(y)\right)$ can be bounded by $2D_1\left(\widetilde{f}(y;\theta^0,\gamma^1)||\widetilde{f}(y;\theta^0,\gamma^0)\right)$ up to terms of order $O\left(||\theta^*-\theta^0||^3+||\gamma^0-\gamma^1||^3\right)$. The advantage of the bound is that it does not require the difficult evaluation of θ^*.

B. Tests Insensitive to a Pitman Drift

Assume again that the parametric class $\{f(y;\theta):\theta\in\Theta\}$ is embedded in a larger class $\{\widetilde{f}(y;\theta,\gamma):\theta\in\Theta,\ \gamma\in\Gamma\subset\mathbb{R}^{K'}\}$ such that $f(y;\theta)=\widetilde{f}(y;\theta,\gamma^0)$ for all $\theta\in\Theta$. Denote by $\beta=[\theta^T,\gamma^T]^T$ the concatenated parameter vector and assume that there exist integrable functions $a(y)$ and $b(y)$ such that $a(y)b(y)$ is integrable as well with respect to ν, and for almost all y, $\widetilde{f}(y;\beta)\leq a(y)$ and $|\log\widetilde{f}(y;\beta)|$, $|\nabla_\beta\log\widetilde{f}(y;\beta)|^2$, $|\nabla^2_\beta\log\widetilde{f}(y;\beta)|$, $|e(y,\theta)|^2$, and $|\nabla_\theta e(y,\theta)|$ are each less than $b(y)$ for all $\beta\in\Theta\times\Gamma$, where for matrices $|\cdot|$ denotes the maximum valued element. Furthermore, assume that the support of $\widetilde{f}(y;\beta)$ in independent of β. Assume that the true underlying distribution depends on n, hence denoted by $g_n(y)$, and is given by

$$g_n(y)=\widetilde{f}(y;\theta^0,\gamma^0+\gamma/\sqrt{n}) \tag{40}$$

for some fixed $\gamma\in\Gamma$, and denote the limiting distribution by $g(y)$. In the context of model specification tests, this type of local alternative is called a Pitman drift. Newey [26] investigated the power of M-tests to such local alternatives. Applying Newey's result to our setting we obtain that if $e(y,\theta)$ satisfies

$$\int e(y,\theta)f(y;\theta)dy=0,\quad \forall\theta\in\Theta$$

then under H_0,

$$\sqrt{n}h_n(\widetilde{\theta}_n)\xrightarrow{D}N\left(D\gamma,V(\theta^0)\right) \tag{41}$$

where in the definition of $V(\theta)$ (12), the expectation is taken with respect to the density $f(y,\theta^0)$ and the term $h(\theta^0)$ vanishes. The term D in (41) is

$$\begin{aligned}D=&\int e(y,\theta^0)\ \nabla^T_\gamma\log\widetilde{f}(y;\theta^0,\gamma)\Big|_{\gamma=\gamma^0}f(y;\theta^0)dy\\ &-H(\theta^0)A^{-1}(\theta^0)\widetilde{B}_{(\theta,\gamma)}(\beta^0)\end{aligned}$$

where the expectations in the definition of $A(\theta)$ and $H(\theta)$, (3) and (11), respectively, are taken with respect to the density $f(y,\theta^0)$ as well. $\beta^0=[\theta^{0T},\gamma^{0T}]^T$ and the matrix $\widetilde{B}_{(\theta,\gamma)}(\beta)$ is the upper right $K\times K'$ block of the FIM associated with the density $\widetilde{f}(y;\beta)$, that is,

$$\widetilde{B}(\beta)=\int\nabla_\beta\log\widetilde{f}(y;\beta)\nabla^T_\beta\log\widetilde{f}(y;\beta)\widetilde{f}(y;\beta)dy, \tag{42}$$

and it is assumed that $\widetilde{B}(\beta)$ is non-singular for all $\beta\in\Theta\times\Gamma$. Hence, S_n, defined in (19), is asymptotically non-central chi-squared distributed with Q degrees of freedom and non-centrality parameter

$$\delta=\gamma' D' V^{-1}(\theta^0)D\gamma.$$

In [26] this result is used to assess and optimize the power of M-tests against local alternatives. Here, our goal is reversed; we would like the tests to be insensitive to small deviations from the assumed model. Specifically, note that

$$\begin{aligned}H(\theta^0)&=\int\nabla_\theta\ e(y,\theta)|_{\theta=\theta^0}f(y;\theta^0)dy\\ &=-\int e(y,\theta)\ \nabla^T_\theta\log\widetilde{f}(y;\theta,\gamma^0)\Big|_{\theta=\theta^0}f(y;\theta^0)dy.\end{aligned}$$

Therefore, considering the space of zero-mean L_2 functions of y with inner product

$$\langle f_1(y),f_2(y)\rangle=\int f_1(y)f_2(y)f(y;\theta)dy$$

our objective is to construct a global maximum validation function $e(y,\theta)$, with elements orthogonal to the space spanned by the $K+K'$ set of functions

$$\nabla_\beta\log\widetilde{f}(y;\beta)\Big|_{\gamma=\gamma^0}. \tag{43}$$

By this construction, both terms of the matrix D are zeroed out, i.e., the test is insensitive to the Pitman drift regardless of the vector γ. Denoting the classes of log-likelihood functions $\{\log f(y;\theta):\theta\in\Theta\}$ and $\{\log\widetilde{f}(y;\theta,\gamma):\theta\in\Theta,\ \gamma\in\Gamma\}$ by $\mathcal{F}$ and $\mathcal{G}$, respectively, Fig. 1 gives a geometrical interpretation of the construction of $e^\perp(y,\theta)$.

Given any global maximum validation function $e(y,\theta)$ that satisfies $\int e(y,\theta)f(y;\theta)dy=0$, $\forall\theta\in\Theta$, its orthogonal component with respect to the vector (43), denoted by $e^\perp(y,\theta)$, is

$$e^\perp(y,\theta)=e(y,\theta)-\left[E(\beta)\widetilde{B}^{-1}(\beta)\nabla_\beta\log\widetilde{f}(y;\beta)\right]_{\gamma=\gamma^0} \tag{44}$$

where $E(\beta)$ is the $K\times(K+K')$ matrix of inner products between the elements of $e(y,\theta)$ and the functions in (43), given by

$$E(\beta)=\int e(y,\theta)\nabla^T_\beta\log\widetilde{f}(y;\beta)f(y;\theta)dy. \tag{45}$$

This can be verified by computing the matrix

$$\int e^\perp(y,\theta)\ \nabla^T_\beta\log\widetilde{f}(y;\beta)\Big|_{\gamma=\gamma^0}f(y;\theta)dy.$$

At any local maximum $\widetilde{\theta}_n$, $\sum_{t=1}^n \nabla_\theta \log f(y_t;\widetilde{\theta}_n) = 0$ and therefore, computing $h_n^\perp(\widetilde{\theta}_n) = \sum_{t=1}^n e^\perp(y_t,\widetilde{\theta}_n)$ reduces to

$$h_n^\perp(\widetilde{\theta}_n) = \sum_{t=1}^n e(y_t,\widetilde{\theta}_n) - E(\beta)\,\widetilde{B}_2(\beta)\sum_{t=1}^n \nabla_\gamma \log \widetilde{f}(y_t;\beta)\Big|_{\theta=\widetilde{\theta}_n,\gamma=\gamma^0}$$

where $\widetilde{B}_2(\beta)$ is the $(K+K')\times K'$ matrix composed of the right K' columns of $\widetilde{B}^{-1}(\beta)$ defined in (42). Furthermore, under the null hypothesis H_0, a consistent estimator for the covariance matrix of $\sqrt{n}h_n^\perp(\widetilde{\theta}_n)$ is

$$\frac{1}{n}\sum_{t=1}^n e^\perp\left(y_t,\widetilde{\theta}_n\right) e^\perp\left(y_t,\widetilde{\theta}_n\right)^T$$

since the term $H(\theta)$ (11), which appears in (28), is zero by construction of $e^\perp(y,\theta)$. When closed form expressions for $E(\beta)$ and $B(\beta)$ are available, the covariance matrix can also be consistently estimated under H_0 by

$$\widetilde{V}_n(\widetilde{\theta}_n) = \int e(y,\widetilde{\theta}_n)e^T(y,\widetilde{\theta}_n)f(y,\widetilde{\theta}_n)dy - E(\widetilde{\theta}_n,\gamma^0)B^{-1}(\widetilde{\theta}_n,\gamma^0)E^T(\widetilde{\theta}_n,\gamma^0). \quad (46)$$

In summary, tests for global maximum which are based on $e^\perp(y,\theta)$ are locally insensitive to model mismatch of the type defined in (40) for any $\gamma\in\Gamma$.

Another motivation for using $e^\perp(y,\theta)$ can be obtained from the Taylor expansion of $h(\theta^*)$ around γ^0. Assuming the derivatives can be taken inside the integrals, we obtain that the zeroth order (constant) term is identically zero and the first order (linear) term is zeroed by the construction of $e^\perp(y,\theta)$.

In practice, we expect these tests to be less sensitive to small deviations from the model. An example in which this is the case is given in Sec. V-A.1.

V. Applications

The asymptotic regime adopted throughout the paper, raises the question of small sample performance. In this section, tests for global maximum will be derived and evaluated through simulations for several parameter estimation problems. In the simulations the following aspects were studied. First, the accuracy of setting the test threshold to $F^{-1}_{\chi^2_Q}(1-\alpha)$ for a level α test was evaluated. Second, we evaluated how fast the power of the test approaches 1, as the number of samples increases, and the accuracy of the finite sample power approximation (37). Finally, the sensitivity of the tests to a misspecified model is examined. The threshold adjustment procedure and the construction of tests that are orthogonal to deviations from the model are demonstrated.

A. Direction Finding in Array Signal Processing

For a review of the problem of direction finding using antenna arrays see e.g. [37] or [38]. The characterization of the MLE under possible model mismatch has been recently addressed in [39] and [35].

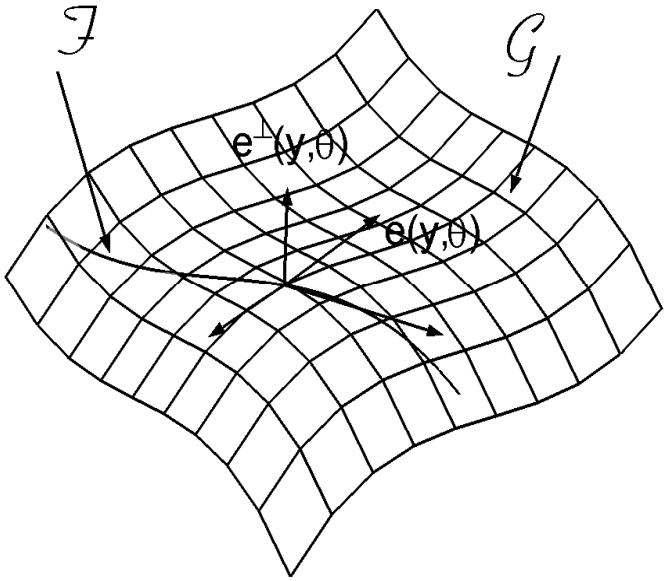

Fig. 1. Geometrical interpretation of the construction of tests insensitive to Pitman drift.

Here we adopt the standard narrow band model of [40]. 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 (λ is the wavelength of wavefronts propagating across the array). The received signal model is given by

$$y_t = D(\theta)s_t + w_t$$

where $y_t\in\mathcal{C}^P$ is the noisy data vector at the array elements,

$$D(\theta) = [d(\theta_1)\quad d(\theta_2)]$$

where $[d(\theta)]_p = \exp\{jp\pi\cos(\theta)\}$, $p=0,1,2,3$ is the steering vector, s_t contains the two signal components, and w_t is a temporally and spatially complex 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 complex circular Gaussian random vector with covariance matrix $C(\theta) = D(\theta)K_sD^H(\theta)+\sigma^2 I$, where, due to an uncorrelated sources assumption, $K_s = \mathrm{diag}(\sigma^2_{s1},\sigma^2_{s2})$, σ^2_{s1} and σ^2_{s2} are the two source variances, and σ^2 is the noise variance. Hence, the density of y is given by

$$f(y,\theta) = \frac{1}{\pi^P\det(C(\theta))}\exp\left[-y^HC^{-1}(\theta)y\right]. \quad (47)$$

The variances σ^2, σ_1^2, and σ_2^2 are assumed known. The only unknowns are the sources directions, $\theta=[\theta_1,\,\theta_2]^T$. In the simulations the true unknown parameters were taken to be $\theta=[\pi/2,\,\pi/2+0.4]^T$ and the other known parameters were set to $\sigma^2_{s1}=\sigma^2_{s2}=1$, and $\sigma^2=2$. In Fig 2, the log-likelihood surface calculated from 200 samples is shown and it is seen that it has two relative maxima.

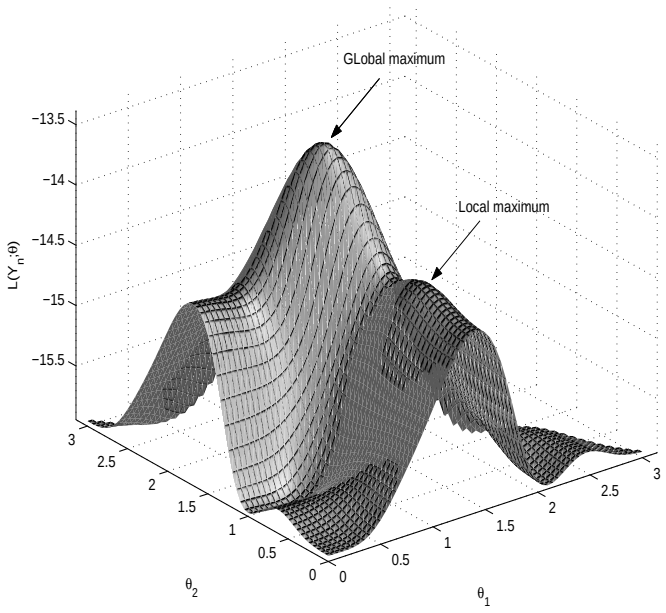

Fig. 2. The log-likelihood function of the direction finding problem.

Recall that the global maximum validation function of Biernacki's test is given by

$$\begin{aligned} e(y,\theta) &= \log f(y;\theta) - \int \log f(y;\theta) f(y;\theta) dy \\ &= -\log\left(\pi^P\right) - \log\left(\det\left(C(\theta)\right)\right) - y^H C^{-1}(\theta) y \\ &\quad + \log\left(\pi^P\right) + \log\left(\det\left(C(\theta)\right)\right) \\ &\quad + \int y^H C^{-1}(\theta) y f(y;\theta) dy \\ &= P - y^H C^{-1}(\theta) y. \end{aligned}$$

Hence

$$\begin{aligned} h_n(\widetilde{\theta}_n) &= \frac{1}{n}\sum_{t=1}^{n} e(y_t, \widetilde{\theta}_n) \\ &= P - \frac{1}{n}\sum_{t=1}^{n} y_t^H C^{-1}(\widetilde{\theta}_n) y_t \\ &= P - \mathrm{tr}\left(C^{-1}(\widetilde{\theta}_n)\widehat{C}\right) \end{aligned}$$

where

$$\widehat{C} = \frac{1}{n}\sum_{t=1}^{n} y_t y_t^H.$$

Under the null hypothesis and assuming the model is correctly specified, a closed form expression for the variance can be computed through (30), where

$$\begin{aligned} \left[\overline{H}(\theta)\right]_{1,i} &= \int \partial e(y,\theta)/\partial\theta_i f(y,\theta) dy \\ &= \int y^H C^{-1}(\theta) \frac{\partial C(\theta)}{\partial \theta_i} C^{-1}(\theta) y f(y,\theta) dy \\ &= \mathrm{tr}\left(C^{-1}(\theta)\frac{\partial C(\theta)}{\partial \theta_i}\right), \quad i = 1,2 \end{aligned}$$

$$\begin{aligned} \int e^2(y,\theta) f(y,\theta) dy &= \int \left[P - y^H C^{-1}(\theta) y\right]^2 f(y,\theta) dy \\ &= P \end{aligned}$$

and $\widetilde{B}(\theta)$ is the FIM for this problem [1, p. 565], given by

$$\left[\overline{B}(\theta)\right]_{i,j} = \mathrm{tr}\left[C^{-1}(\theta)\frac{\partial C(\theta)}{\partial \theta_i} C^{-1}(\theta)\frac{\partial C(\theta)}{\partial \theta_j}\right]. \qquad (48)$$

Hence

$$\overline{V}(\widetilde{\theta}_n) = P - \overline{H}(\widetilde{\theta}_n)\overline{B}(\widetilde{\theta}_n)\overline{H}^T(\widetilde{\theta}_n)$$

and the test statistic is given by

$$S_n = n\left[P - \mathrm{tr}\left(C^{-1}(\widetilde{\theta}_n)\widehat{C}\right)\right]^2 / \overline{V}(\widetilde{\theta}_n). \qquad (49)$$

The threshold is set according to a χ^2 distribution with one degree of freedom.

We compare Biernacki's test to a test which is based on the real part of the first off-diagonal element of the covariance matrix. To compare the first off-diagonal element of the covariance matrix at the candidate relative maximum to its unconstrained estimate from the data, the global maximum validation function is taken to be

$$e(y,\theta) = y^H M y - \mathrm{tr}\left(MC(\theta)\right)$$

where M is the symmetric Toeplitz matrix whose first row is $[0,1,0,0]$, and hence

$$\begin{aligned} h_n(\widetilde{\theta}_n) &= \frac{1}{n}\sum_{t=1}^{n} e(y_t, \widetilde{\theta}_n) \\ &= \mathrm{tr}\left(M\widehat{C}\right) - \mathrm{tr}\left(MC(\widetilde{\theta}_n)\right). \end{aligned}$$

For this choice of $e(y,\theta)$ we have

$$\left[\overline{H}(\theta)\right]_{1,i} = -\mathrm{tr}\left(M\frac{\partial C(\theta)}{\partial \theta_i}\right), \quad i = 1,2 \qquad (50)$$

and by [1, p. 564]

$$\begin{aligned} \int e^2(y,\theta) f(y,\theta) dy &= \\ &\int \left[y^H M y - \mathrm{tr}\left(MC(\theta)\right)\right]^2 f(y,\theta) dy \\ &= \mathrm{tr}\left(MC(\theta)MC(\theta)\right). \end{aligned}$$

Hence

$$\overline{V}(\widetilde{\theta}_n) = \mathrm{tr}\left(MC(\widetilde{\theta}_n)MC(\widetilde{\theta}_n)\right) - \overline{H}(\widetilde{\theta}_n)\overline{B}(\widetilde{\theta}_n)\overline{H}^T(\widetilde{\theta}_n)$$

the test statistic is given by

$$S_n = n\left[\mathrm{tr}\left(M\widehat{C}\right) - \mathrm{tr}\left(MC(\widetilde{\theta}_n)\right)\right]^2 / \overline{V}(\widetilde{\theta}_n) \qquad (51)$$

and, again, the threshold is set according to a χ^2 distribution with one degree of freedom.

The power performance of Biernacki's test and a Covariance based test were evaluated for increasing n for levels that were set to 0.01 and 0.001. 1000 Monte Carlo iterations were used. At each iteration the global maximum and the local maximum were found and the tests were applied to both maxima to evaluate the performance. When the number of samples is very

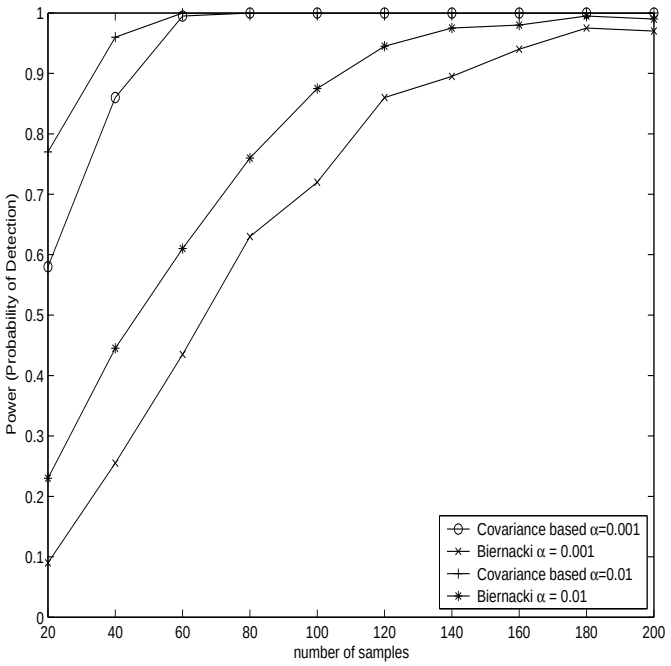

Fig. 3. Direction finding: power when the model is correctly specified.

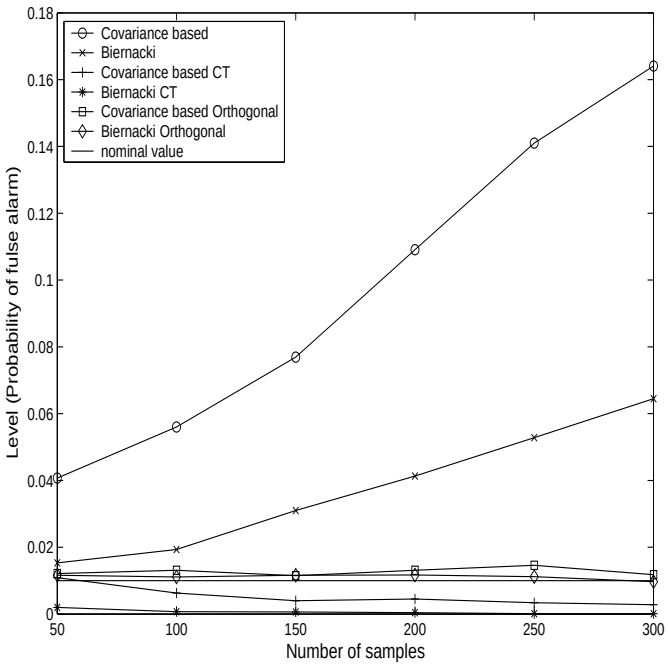

Fig. 4. Direction finding: level under model mismatch.

small (e.g. $n = 20$), the likelihood function my be distorted and the two relative maxima may collapse into one. Such cases were eliminated from the analysis. The results are summarized in Fig. 3. While not presented here, we observed that the empirical levels of both tests were in good agreement with the specified values.

1) Model Mismatch: In this section the performance of the tests (49) and (51) under model mismatch is evaluated. The assumed model used for the estimation is the same as in the previous section (47). The samples were generated according to the model (47) but with covariance matrix

$$C(\theta,\gamma) = D(\theta)K_s D^H(\theta) + \sigma^2 R(\gamma), \tag{52}$$

where $R(\gamma)$ is a symmetric Toeplitz matrix whose first row is $[1,\gamma,\gamma^2,\gamma^3]$, which corresponds to a first order AR spatial noise covariance [41], and in the simulation $\gamma = 0.1$.

For both Biernacki's test and the covariance based test the effect of model mismatch on the level was evaluated for three cases: (a) The increase in level due to model mismatch when the tests are performed without any adjustment, (b) The threshold correction described in Sec. IV-A, and (c) The performance of the orthogonal counterparts given in Sec. IV-B.

To perform the threshold correction described in Sec. IV-A, the Kullback-Leibler distance needs to be estimated. In the simulation, it was assumed that it is known that the parameter γ, which controls the deviation from the model, ranges between zero (correct model) and 0.1. At every Monte Carlo iteration, given a relative maximum $\widetilde{\theta}_n$,

$$c = \max_{\gamma\in[0,1]} D_1\left(\widetilde{f}(y;\widetilde{\theta}_n,\gamma)||f(y;\widetilde{\theta}_n)\right)$$

was computed, using the known formula for the Kullback-Leibler distance between two Gaussian densities (e.g. [42]), where $f(y;\theta)$ is given in (47) and $\widetilde{f}(y;\theta,\gamma)$ is the same density but with covariance matrix $C(\theta,\gamma)$ (52). Then, the null hypothesis was rejected if

$$S_n > F^{-1}_{\chi^2_Q(n[\exp(2c)-1])}(1-\alpha).$$

The simulation results show that, as anticipated, the level decreases rather than increases with the number of samples (see Fig. 4, where CT is a shorthand notation for 'corrected threshold').

To construct the orthogonal counterparts of the two tests, $e^{\perp}(y,\theta)$ is found through (44). For Biernacki's test the elements of $E(\beta)$ (45), which is a 1×3 vector is this case, are given by

$$[E(\beta)]_i = -\mathrm{tr}\left(C^{-1}(\beta)\frac{\partial C^{-1}(\beta)}{\partial\beta_i}\right), \quad i=1,2,3$$

where, as defined earlier, $\beta = [\theta^T,\gamma]^T$. For the covariance based test the elements of $E(\beta)$ are given by

$$[E(\beta)]_i = \mathrm{tr}\left(M\frac{\partial C^{-1}(\beta)}{\partial\beta_i}\right), \quad i=1,2,3.$$

The FIM $\widetilde{B}(\beta)$ is also available in closed form as given in (48). Using the closed forms for $E(\beta)$ and $\widetilde{B}(\beta)$, the variance for the two tests was computed through (46). In Fig. 4 it is seen that while the original tests suffer from increased level as the number of samples increase, the orthogonal counterparts are unaffected by this type of model mismatch.

B. Estimation of Gaussian Mixture Parameters

The problem of estimation of Gaussian mixture parameters arises in both non-parametric density estimation (see e.g. [43] and references therein) and a variety of clustering problems

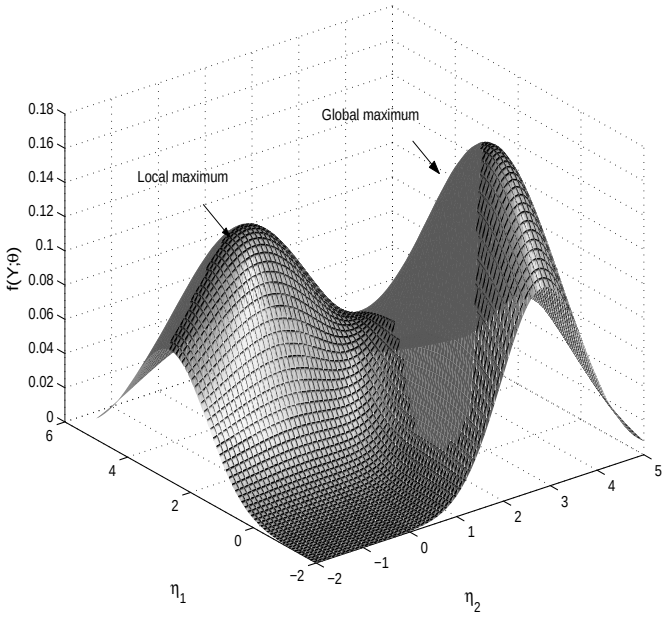

Fig. 5. The likelihood function of the Gaussian mixture distribution.

(see e.g. [44] and references therein). The MLE for this problem is usually found by using the EM algorithm [10]. In [44], the authors describe a method that finds the global maximum with good performance. However, even this state of the art method is not certain to find the global maximum, and therefore, tests for global maximum are useful.

Here we consider the univariate case, in which the independent scalar measurements are generated by the following two component univariate Gaussian mixture density

$$f(y;\theta)=\sum_{l=1}^{2}\frac{p_l}{\sqrt{2\pi\sigma_l^2}}\exp\left\{-\frac{(y-\eta_l)^2}{2\sigma_l^2}\right\} \qquad (53)$$

where the parameter vector consists of the two means $\theta = [\eta_1\ \eta_2]^T$. The number of components, the variances, and the mixing probabilities are assumed known. In the simulation, the true parameter is $\theta = [0,\ 3]^T$, the variances are $\sigma_1^2 = 1$ and $\sigma_1^2 = 0.5$, the mixing probabilities are $p_1 = 1-p_2 = 0.35$ and it is known that $\Theta = [-1,\ 4]\times[-1,\ 4]$. The likelihood surface over Θ of a realization of 200 samples generated according to this model is presented in Fig. 5 and two relative maxima appear.

The performance of the global maximum tests was evaluated as the number of samples n increases. 1000 Monte Carlo iterations were generated. At each iteration, Biernacki's test and a mean based test were performed on both the global and the local maxima. As in the previous section, Biernacki's global maximum validation function is given by

$$e(y,\theta)=\log f(y;\theta)-\int \log f(y;\theta)f(y;\theta)dy \qquad (54)$$

and therefore,

$$h_n(\widetilde{\theta}_n)=\frac{1}{n}\sum_{t=1}^{n}\log f(y_t;\widetilde{\theta}_n)-\int \log f(y;\widetilde{\theta}_n)f(y;\widetilde{\theta}_n)dy.$$

A closed form expression to the integral in (54) is not available. Hence, in the simulations, numerical integration is used. The variance $V_n(\widetilde{\theta}_n)$ required for the construction of the test statistic S_n (19) was calculated through (13). Note that $H_n(\theta)$, required for calculating $V_n(\widetilde{\theta}_n)$, simplifies under the null hypothesis, i.e. $\widetilde{\theta}_n = \widehat{\theta}_n$, to

$$\begin{aligned}H_n(\widetilde{\theta}_n) &= \frac{1}{n}\sum_{t=1}^{n}\nabla_\theta^T e(y_t,\theta)\Big|_{\theta=\widetilde{\theta}_n}\\ &= \frac{1}{n}\sum_{t=1}^{n}\nabla_\theta^T \log f(y;\theta)\\ &\quad -\int \nabla_\theta^T \log f(y;\theta)f(y;\theta)dy\\ &\quad -\int \log f(y;\theta)\nabla_\theta^T f(y;\theta)dy\Big|_{\theta=\widetilde{\theta}_n}\\ &= -\int \log f(y;\theta)\nabla_\theta^T f(y;\theta)dy\Big|_{\theta=\widetilde{\theta}_n}\end{aligned}$$

which was calculated in the simulation by numerical integration.

The global maximum validation function of the mean based test is given by

$$e(y,\theta)=y-[p\eta_1+(1-p)\eta_2]$$

which leads to

$$h_n(\widetilde{\theta}_n)=\frac{1}{n}\sum_{t=1}^{n}y_t-(p\widetilde{\eta}_1+(1-p)\widetilde{\eta}_2). \qquad (55)$$

Similar to the previous test, the variance required for the test statistic was calculated through (13), where, for this test, the vector $H_n(\widetilde{\theta}_n)$ is given by

$$H_n(\widetilde{\theta}_n)=-\left[p,\ (1-p)\right].$$

The level of the tests was set to 0.01 and the empirical power was estimated from $10,000$ Monte Carlo iterations and compared to the analytic approximation (37). The results are summarized in Fig. 6 and it can be seen that the analytical power approximation predicts the empirical power well. It can be seen that the power of the mean based test is better than that of Biernacki's test. For other choices of parameters different results may be obtained. While not reported here, the empirical level of both tests was in good agreement with its specified value.

C. Estimation of Superimposed Exponentials in Noise

For a review of the problem of estimating the parameters of superimposed exponentials in noise see, e.g., [40]. Consider the following model

$$y_t=\sum_{k=1}^{K}\alpha_k\exp\{j\Omega_k t\}+w_t,\quad t=1,\ldots,n$$

where w_t is a white circular Gaussian noise with unknown variance σ^2. The unknown parameters are the frequencies of the exponentials $[\Omega_1,\ldots,\Omega_K]$, their complex valued amplitudes $[\alpha_1,\ldots,\alpha_K]$ and the noise variance. The number K

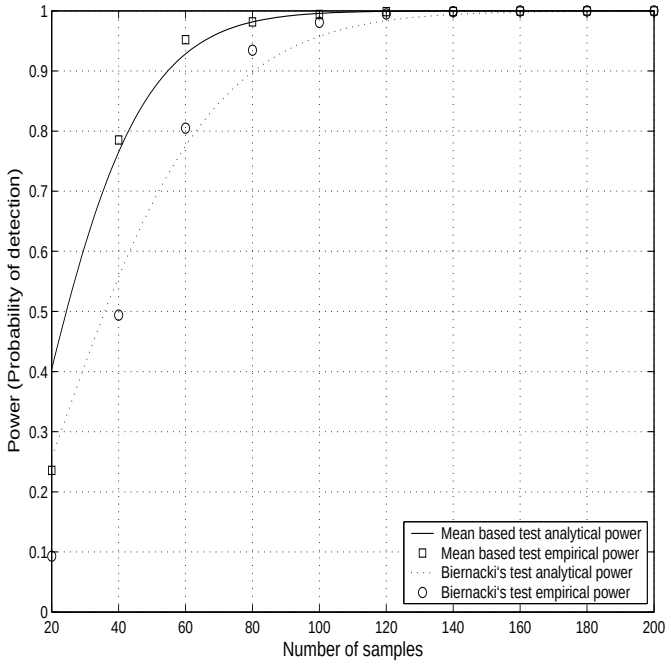

Fig. 6. Gaussian mixture: empirical power vs. its analytic prediction, when the level is set to 0.01.

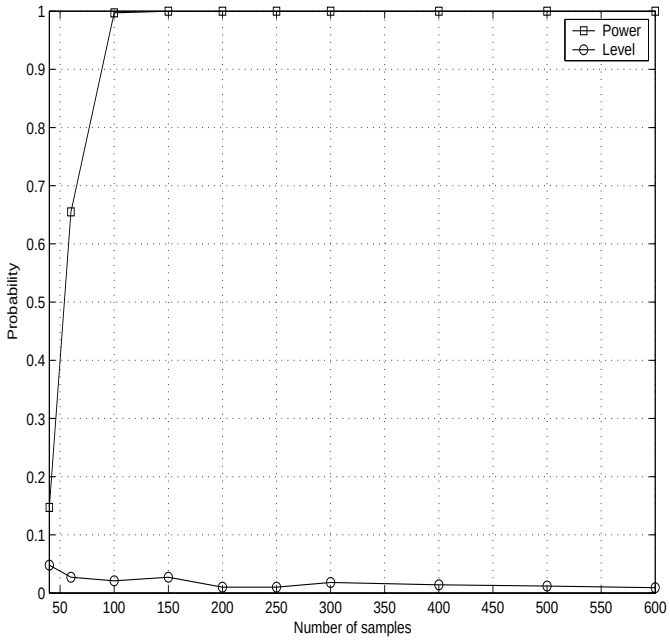

Fig. 7. Exponentials in noise: performance when the model is correctly specified.

of components is assumed known and was set to 3, hence there are 10 unknown parameters. The unknown parameters were set to $[\Omega_1, \Omega_2, \Omega_3] = [0.4,\, 0.5,\, 0.6]$, $[\alpha_1, \alpha_2, \alpha_3] = [\exp(j2),\, 0.8\exp(j3),\, 1.2\exp(j5)]$, and $\sigma^2 = 1$.

Under this generating model, the data are independent but not identically distributed. They are distributed as non-zero time-varying mean circular Gaussian process. Hence, the treatment in Sec. II-A does not cover this problem. Furthermore, since the MLE for this problem is super efficient [45], the more general framework of White [29] for constructing tests in dynamical models does not cover this problem either. However, a detailed statistical asymptotic analysis for this problem is available in the literature and can be used to construct a test for global maximum. In particular, in [45] it was shown that the MLE is asymptotically normal distributed under an appropriate normalization. Based on this analysis, we propose a test which is based on the autocorrelation function. In particular, our test is based on the fact that at the true parameter,

$$\mathrm{E}\left\{\left[y_t - \sum_{k=1}^{K}\alpha_k \exp(j\Omega_k t)\right] \times \left[y_{t-1} - \sum_{k=1}^{K}\alpha_k \exp(j\Omega_k (t-1))\right]^*\right\} = \mathrm{E}\left\{e_t e_{t-1}^*\right\} = 0,$$

and hence, given the local maximum $\widetilde{\theta}_n$, we construct a test from the real part of the statistic

$$h_n(\widetilde{\theta}_n) = \frac{1}{n-1}\sum_{t=2}^{n}\left[y_t - \sum_{k=1}^{K}\widetilde{\alpha}_k \exp(j\widetilde{\Omega}_k t)\right] \times \left[y_{t-1} - \sum_{k=1}^{K}\widetilde{\alpha}_k \exp(j\widetilde{\Omega}_k (t-1))\right]^*.$$

It is shown in the Appendix that under the null hypothesis, the real part of this statistic is asymptotically distributed as a zero-mean Gaussian random variable with variance $\sigma^2/2$. Hence, since under the null hypothesis $\widetilde{\sigma}^2$ is a consistent estimator for σ^2, the statistic

$$n\frac{\left(\Re\{h_n(\widetilde{\theta}_n)\}\right)^2}{\widetilde{\sigma}^2/2}$$

is asymptotically χ^2 distributed with one degree of freedom, and can be used to discriminate between local and global maxima. In Fig. 7 the performance of this test is presented when the level is set to 0.01. The empirical level and power of the test were estimated from 1000 Monte Carlo iterations. It is seen that the asymptotic approximation to the level α is accurate for n greater then 300 and the power of the test approaches 1 when n is greater then 100.

VI. Concluding Remarks

This paper has investigated 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 for exploring solutions 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

practical implementation of this tool. 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 $e(y,\theta)$ in terms of achieving maximum power for a given level with minimum sensitivity to model mismatch? This remains an open question.

It is possible to generalize the above concept to non-i.i.d. measurements. A unified treatment of the MLE under a possible model mismatch and the construction of model mismatch tests for dynamic models is given in [29] and an example is which the measurements are i.n.i.d. was treated in Sec. V-C. The concept of using a statistical test for discriminating between global and local maxima can be generalized to other M-estimators [2], or any other optimization problem in which a statistical characterization of the global maximum is available.

APPENDIX I
ASYMPTOTIC DISTRIBUTION OF M-TESTS

The proof follows White's methodology [29]. Given the assumptions, the mean value theorem for random functions, given as Lemma 3 in [33], guarantees the existence of measurable Θ-valued functions $\overline{\theta}_n$ such that

$$\sqrt{n}h_n(\widehat{\theta}_n) = \sqrt{n}h_n(\theta^*) + H_n(\overline{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \quad \text{(I.56)}$$

where each $\overline{\theta}_n$ lies on the segment joining $\widehat{\theta}_n$ and θ^*. Each row of H_n depends on a different $\overline{\theta}_n$, but since it makes no difference asymptotically, the above shorthand notation is used. From (5) $\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right)$ converges in distribution. Furthermore, $\widehat{\theta}_n \stackrel{a.s.}{\to} \theta^*$ and therefore $\overline{\theta}_n \stackrel{a.s.}{\to} \theta^*$ as well. From Theorem 2 in [33], applied on the elements of $H_n(\theta)$, we have $H_n(\theta) \stackrel{a.s.}{\to} H(\theta)$ uniformly in θ, and therefore using Lemma 3.1 of White [?], $H_n(\overline{\theta}_n) - H(\theta^*) \stackrel{a.s.}{\to} 0$. Using these intermediate results we obtain from $2c.4(xa)$ of Rao [46] that

$$\left[H_n(\overline{\theta}_n) - H(\theta^*)\right]\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \stackrel{P}{\to} 0. \quad \text{(I.57)}$$

Equation (A.2) of [20] asserts that

$$A^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla \log f(y_t,\theta^*) + \sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \stackrel{P}{\to} 0.$$

Therefore, by the finiteness of $H(\theta^*)$, we have

$$H(\theta^*) \times \left[A^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla \log f(y_t,\theta^*) + \sqrt{n}\left(\widehat{\theta}_n - \theta^*\right)\right] \stackrel{P}{\to} 0.$$

Adding and subtracting $H_n(\overline{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right)$ and rearranging terms, we obtain

$$\left[H(\theta^*) - H_n(\overline{\theta}_n)\right]\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) + H_n(\overline{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) + H(\theta^*)A^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla \log f(y_t,\theta^*) \stackrel{P}{\to} 0.$$

But from (I.57) the first term converges to zero in probability, and hence,

$$H_n(\overline{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) + H(\theta^*)A^{-1}(\theta^*)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla \log f(y_t,\theta^*) \stackrel{P}{\to} 0.$$

Substituting $H_n(\overline{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) = \sqrt{n}h_n(\widehat{\theta}_n) - \sqrt{n}h_n(\theta^*)$ from (I.56), adding and subtracting $\sqrt{n}h(\theta^*)$, and rearranging terms, we obtain

$$\sqrt{n}\left[h_n(\widehat{\theta}_n) - h(\theta^*)\right] - \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[e(y_t,\theta^*) - h(\theta^*) - H(\theta^*)A^{-1}(\theta^*)\nabla \log f(y_t,\theta^*)\right] \stackrel{P}{\to} 0.$$

From the Lindeberg-Lévy central limit theorem the second term converges in probability to a zero mean multivariate normal density, with covariance matrix $V(\theta^*)$ and therefore, from $2c.4(xd)$ of Rao [46], so does the first term, and the first part of the theorem is proved. The consistency of $V_n(\widehat{\theta}_n)$ for $V(\theta^*)$ follows from Lemma 3.1 of White [?] given the assumptions, and the consistency guarantees that $V_n^{-1}(\widehat{\theta}_n)$ exists for sufficiently large n, since the determinant of a matrix is a continuous function of its elements. The last part of the theorem follows from Lemma 3.3 of White [47] and the proof is completed.

APPENDIX II
ASYMPTOTIC DISTRIBUTION OF THE TEST STATISTIC FOR EXPONENTIALS IN NOISE

The derivation is given under the null hypothesis, hence $\widetilde{\theta}_n = \widehat{\theta}_n$. Using the mean value theorem we obtain

$$h_n(\widehat{\theta}_n) = h_n(\theta^0) + \nabla^T h_n(\overline{\theta})(\widehat{\theta}_n - \theta^0) \quad a.s..$$

Using the martingale central limit theorem with the filtration $\{\mathcal{F}_t = \sigma(e_1,\ldots,e_t)\}$ [48], we obtain that $h_n(\theta^0)$ converges in distribution to a zero-mean Gaussian random variable with variance $\sigma^2/2$. Next, we show that the second term is $\mathrm{o}_P(1)$. First split the second term into two components

$$\nabla^T h_n(\overline{\theta})(\widehat{\theta}_n - \theta^0) = n^{-3/2}\nabla_\Omega^T h_n(\overline{\theta})n^{3/2}(\widehat{\mathbf{\Omega}}_n - \mathbf{\Omega}^0) + n^{-1/2}\nabla_\alpha^T h_n(\overline{\theta})n^{1/2}(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^0).$$

It is possible to show that both $n^{-3/2}\nabla_\Omega^T h_n(\overline{\theta})$ and $n^{-1/2}\nabla_\alpha^T h_n(\overline{\theta})$ converge to zero in probability. Therefore, since it was shown in [45] that both $n^{3/2}(\widehat{\mathbf{\Omega}}_n - \mathbf{\Omega}^0)$ and $n^{1/2}(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^0)$ converge in distribution, we have that this term converges to zero in probability. This establish the asymptotic normality of $h_n(\widehat{\theta}_n)$. In [45] it was also shown that $\widehat{\sigma}^2$ converges to the true value of σ^2 a.s.. Therefore, by Lemma 3.3 of White [47], we obtain that the test statistic is asymptotically χ^2 distributed.

REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper saddle river, New Jersey: Prentice Hall, 1993.
[2] P. Huber, *Robust Statistics*. New York: John Wiley & Sons, 1981.

[3] H. L. V. Trees, *Detection, Estimation, and Modulation Theory*. New York: John Wiley & Sons, 2001.
[4] C. Andrieu and A. Doucet, "Simulated annealing for maximum *a posteriori* parameter estimation of hidden markov models," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 994 – 1004, May 2000.
[5] C. Bon-Sen, L. Bore-Kuen, and P. Sen-Chueh, "Maximum likelihood parameter estimation of F-ARIMA processes using the genetic algorithm in the frequency domain," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2208 – 2220, Sept. 2002.
[6] C. H. Slump and B. J. Hoenders, "The determination of the location of the global maximum of a function in the presence of several local extrema," *IEEE Trans. Inform. Theory*, vol. 31, no. 4, pp. 490 – 497, July 1985.
[7] S. F. Yau and Y. Bresler, "Maximum likelihood parameter estimation of superimposed signals by dynamic programming," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 804–820, Feb. 1993.
[8] I. Ziskind and M. Wax, "Maximum likelihood localization of diversely polarized sources by simulated annealing," *IEEE Trans. Antennas Propagat.*, vol. 38, no. 7, pp. 1111–1114, July 1990.
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data using the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[10] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 1997.
[11] D. Starer and A. Nehorai, "Newton algorithms for conditional and unconditional maximum likelihood estimation of the parameters of exponential signals in noise," *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1528–1534, June 1992.
[12] H. Erdogan and J. A. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 801–814, Sept. 1999.
[13] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[14] S. J. Finch, N. R. Mendell, and H. C. Thode, Jr., "Probabilistic measure of adequacy of a numerical search for a global maximum," *J. Amer. Statist. Assoc.*, vol. 84, no. 408, pp. 1020–1023, Dec. 1989.
[15] P. J. Bickel and J. A. Yahav, "On estimating the total probability of the unobserved outcomes of an experiment," in *Adaptive statistical procedures and related topics*, J. van Ryzin, Ed. Hayward, CA: Institute of Mathematical Statistics 1986.
[16] M. R. Veall, "Testing for a glomal maximum in an econometric context," *Econometrica*, vol. 58, no. 6, pp. 1459–1465, Nov. 1990.
[17] L. de Haan, "Estimation of the minimum of a function using order statistics," *J. Amer. Statist. Assoc.*, vol. 76, no. 374, pp. 467–469, June 1981.
[18] R. E. Dorsey and W. J. Mayer, "Detection of spurious maxima through random draw tests and specification tests," *Computational Economics*, vol. 16, pp. 237–256, 2000.
[19] L. Gan and J. Jiang, "A test for global maximum," *J. Amer. Statist. Assoc.*, vol. 94, no. 447, pp. 847–854, Sept. 1999.
[20] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–26, Jan. 1982.
[21] C. Biernacki, "Un test pour le maximum global de vraisemblance," in *35ièmes journées de statistiques, SFdS'2003*, Lyon, France, June 2003.
[22] ——, "Testing for a global maximum of the likelihood," *J. Comput. Graph. Statist.*, 2004, in press. [Online]. Available: http://www-math.univ-fcomte.fr/pp_Annu/CBIERNACKI/testML_full_version.pdf
[23] D. R. Cox, "Tests of separate families of hypotheses," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press, 1961, pp. 105–123.
[24] ——, "Further results on tests of separate families of hypotheses," *J. Roy. Statist. Soc. Ser. B*, vol. 24, no. 2, pp. 406–424, 1962.
[25] C. G. Small, J. Wang, and Z. Yang, "Eliminating multiple root problems in estimation," *Statist. Sci.*, vol. 15, no. 4, pp. 313 – 341, 2000.
[26] W. K. Newey, "Maximum likelihood specification testing and conditional moment tests," *Econometrica*, vol. 55, pp. 1047–1070, 1985.
[27] G. Tauchen, "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, vol. 30, pp. 415–444, 1985.
[28] H. White, "Specification testing in dynamic models," in *Advances in Econometrics*, T. Bewley, Ed. New York: Cambridge University Press, 1987.
[29] ——, *Estimation, Inference and Specification Analysis*. Cambridge University Press, 1994.
[30] ——, "Consequences and detection of misspecified nonlinear regression models," *J. Amer. Statist. Assoc.*, vol. 76, no. 374, pp. 419–433, Jun. 1981.
[31] W. K. Newey and K. D. West, "Automatic lag selection in covariance matrix estimation," *Reviews of Econometric Studies*, vol. 61, pp. 631–653, 1994.
[32] H. Bunzel, N. M. Kiefer, and T. J. Vogelsang, "Simple robust testing of hypotheses in nonlinear models," *J. Amer. Statist. Assoc.*, vol. 96, no. 455, pp. 1088–1096, Sept. 2001.
[33] R. I. Jennrich, "Asymptotic properties of non-linear least squares estimators," *Ann. Math. Statist.*, vol. 40, no. 2, pp. 633–643, Apr. 1969.
[34] N. L. Johnson, S. Kotz, and A. Balkrishnan, *Continuous univariate distributions: Vol. 2*. Wiley, New York, 1994.
[35] W. Xu, A. B. Baggeroer, and K. L. Bell, "A bound on mean-square estimation error with background parameter mismatch," *IEEE Trans. Inform. Theory*, vol. 50, no. 4, pp. 621–632, Apr. 2004.
[36] S. Amari, *Differential-Geometrical Methods in Statistics*. Berlin: Springer-Verlag, 1990.
[37] A. Hero *et al.*, "Highlights of statistical signal and array processing," *IEEE Signal Processing Mag.*, vol. 15, no. 5, pp. 21–64, Sept. 1998.
[38] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Mag.*, vol. 13, no. 4, pp. 67–94, July 1996.
[39] J. Friedmann, E. Fishler, and H. Messer, "General asymptotic analysis of the generalized likelihood ratio test for a Gaussian point source under statistical or spatial mismodeling," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2617–2631, Nov. 2002.
[40] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramér Rao bound," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 5, pp. 720–741, May 1989.
[41] V. Nagesha and S. Kay, "Maximum likelihood estimation for array processing in colored noise," *IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 169–180, Feb. 1996.
[42] B. C. Levy and R. Nikoukhah, "Robust least-squares estimation with a relative entropy constraint," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 89–104, Jan. 2004.
[43] R. D. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Trans. Signal Processing*, vol. 51, no. 8, pp. 2245 – 2253, Aug. 2003.
[44] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.
[45] C. R. Rao and L. C. Zhao, "Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals," *IEEE Trans. Signal Processing*, vol. 41, no. 3, pp. 1461–1464, Mar. 1993.
[46] C. R. Rao, *Linear Statistical Inference and Its Applications*. John Wiley & Sons, 1973.
[47] H. White, "Nonlinear regression on cross-section data," *Econometrica*, vol. 48, no. 3, pp. 721–746, Apr. 1980.
[48] P. Billingsley, *Probability and Measure*. New York: John Wiley and Sons, 1995.