Asymptotic Characterization of Log-Likelihood Maximization Based Algorithms and Applications

Doron Blatt and Alfred Hero

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
dblatt@umich.edu, hero@eecs.umich.edu

Abstract. The asymptotic distribution of estimates that are based on a sub-optimal search for the maximum of the log-likelihood function is considered. In particular, estimation schemes that are based on a two-stage approach, in which an initial estimate is used as the starting point of a subsequent local maximization, are analyzed. We show that asymptotically the local estimates follow a Gaussian mixture distribution, where the mixture components correspond to the modes of the likelihood function. The analysis is relevant for cases where the log-likelihood function is known to have local maxima in addition to the global maximum, and there is no available method that is guaranteed to provide an estimate within the attraction region of the global maximum. Two applications of the analytic results are offered. The first application is an algorithm for finding the maximum likelihood estimator. The algorithm is best suited for scenarios in which the likelihood equations do not have a closed form solution, the iterative search is computationally cumbersome and highly dependent on the data length, and there is a risk of convergence to a local maximum. The second application is a scheme for aggregation of local estimates, e.g. generated by a network of sensors, at a fusion center. This scheme provides the means to intelligently combine estimates from remote sensors, where bandwidth constraints do not allow access to the complete set of data. The result on the asymptotic distribution is validated and the performance of the proposed algorithms is evaluated by computer simulations.

Keywords – Maximum likelihood, mixture models, clustering, sensor networks, data fusion.

1 Introduction

The maximum likelihood (ML) estimation method introduced by Fisher [1] is one of the standard tools of statistics. Among its appealing properties are consistency and asymptotic efficiency [2]. Furthermore, its asymptotic Gaussian distribution makes the asymptotic performance analysis tractable [2]. However, one drawback of this method is the fact that the associated likelihood equations

required for the derivation of the estimator rarely have a closed form analytic solution. Therefore, suboptimal iterative maximization procedures are used. In many cases, the performance of these methods depends on the starting point. In particular, if the likelihood function of a specific statistical model does not have a known strictly convex property and there is no available method that is guaranteed to provide a starting point within the attraction region of the global maximum, then there is a risk of convergence to a local maximum, which leads to large-scale estimation errors.

The first part of this paper considers the asymptotic distribution of estimates that are based on a sub-optimal search for the ML estimate. In particular, estimators that are based on a two-stage approach, in which an initial estimate is used as the starting point of a subsequent iterative search that converges to a maximum point, are analyzed and shown to be asymptotically Gaussian mixture distributed. The results are linked to previous results by Huber [3], White [4], and Gan and Jiang [5] as explained in detail below.

In the second part of the paper, two applications of the analytical results are presented. The first is an algorithm for finding the ML estimate. The algorithm is best suited for scenarios in which the likelihood equations do not have a closed form solution, the iterative search is computationally cumbersome and highly dependent on the data length, and there is a risk of convergence to a local maximum. The algorithm is performed in two stages. In the first stage, the data are divided into sub-blocks in order to reduce the computational burden, and local estimates are computed from each block. The second stage involves clustering of these local estimates using a finite Gaussian mixture model, which is a classic problem in statistical pattern recognition (e.g. [6], [7], and references therein.) The second application arises in distributed sensor networks. In particular, consider a case where a large number of sensors are distributed in order to perform an estimation task. Due to power and bandwidth constraints the sensors do not transmit the complete data but rather only a suboptimal estimate. As will be shown, the analytical results provide the means for combining these sub-optimal estimates into a final estimate.

2 Problem Formulation

The independent random vectors $\mathbf{y}_n$, $n = 1, \ldots, N$ have a common probability density function (p.d.f.) $f(\mathbf{y}; \boldsymbol{\theta})$, which is known up to a vector of parameters $\boldsymbol{\theta} = [\theta_1 \theta_2 \ldots \theta_K]^T \in \boldsymbol{\Theta}$. The unknown true parameter vector will be denoted by $\boldsymbol{\theta}^0$. The log-likelihood of the measurements under $f(\mathbf{y}; \boldsymbol{\theta})$ is

$$L_N(\mathbf{Y}; \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln f(\mathbf{y}_n; \boldsymbol{\theta}) \ , \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1 \, \mathbf{y}_2 \, \ldots \, \mathbf{y}_N]$. The ML estimator (MLE) for $\boldsymbol{\theta}$, which will be denoted by $\widehat{\boldsymbol{\theta}}_N$ is

$$\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} L_N(\mathbf{Y}; \boldsymbol{\theta}) \ . \tag{2}$$

In many cases, the above maximization problem does not have an analytical solution, and a sub-optimal maximization technique is used. One possible method could be the following. First, a sub-optimal algorithm generates a rough estimate for $\boldsymbol{\theta}$. Then, this rough estimate is used as the starting point of an iterative algorithm, which searches for the maximum of the log-likelihood function. Among those are the standard maximum search algorithms, such as the steepest ascent method, Newton's algorithm, the Nelder-Mead method, and the statistically derived expectation maximization algorithm [8] and its variations. This class of methods will be referred to as two-stage methods, and the resulting estimator will be denoted by $\widetilde{\boldsymbol{\theta}}_N$. If the starting point of the search algorithm is within the attraction region of the global maximum (with respect to the specific searching technique), then this approach leads to the MLE. However, if the likelihood function has more than one maximum and if the staring point is not within the attraction region of the global maximum, then the algorithm will converge to a local maximum resulting in a large-scale estimation error. In the next section, the asymptotic p.d.f. of $\widetilde{\boldsymbol{\theta}}_N$ is derived. The derivation is performed using conditional distributions, where the conditioning is on the location of the initial estimator in $\boldsymbol{\Theta}$.

3 Asymptotic Analysis

The maximization of $L_N(\mathbf{Y};\boldsymbol{\theta})$ is identical to the maximization of $\frac{1}{N}L_N(\mathbf{Y};\boldsymbol{\theta})$, which, due to the law of large numbers, converges almost surely (a.s.) to the ambiguity function

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\ln f(\mathbf{y}_n;\boldsymbol{\theta}) &\rightarrow \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\ln f(\mathbf{y};\boldsymbol{\theta})\right\} \quad \text{a.s.} \\ &= \int_{\mathcal{Y}} \ln\left(f(\mathbf{y};\boldsymbol{\theta})\right) f(\mathbf{y};\boldsymbol{\theta}^0)d\mathbf{y} \stackrel{\triangle}{=} g(\boldsymbol{\theta}^0,\boldsymbol{\theta}) \ , \end{aligned} \tag{3}$$

where $\mathrm{E}_{\boldsymbol{\theta}^0}\{\cdot\}$ denotes the statistical expectation with respect to the true parameter $\boldsymbol{\theta}^0$, and $\mathrm{E}_{\boldsymbol{\theta}^0}\{\ln f(\mathbf{y};\boldsymbol{\theta})\}$ is assumed to be finite for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Therefore, asymptotically, the two-stage method will result in an estimate which is in the vicinity of one of the local maxima of the ambiguity function. The ambiguity function has its global maximum at the true parameter $\boldsymbol{\theta}^0$ [9], and it is assumed to have a number of local maxima in $\boldsymbol{\Theta}$ at points which will be denoted by $\boldsymbol{\theta}^m$, $m=1,\ldots,M$. All the local maxima satisfy

$$\left.\frac{\partial g(\boldsymbol{\theta}^0,\boldsymbol{\theta})}{\partial\theta_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^m} = 0, \quad m=0,\ldots,M, \quad k=0,\ldots,K \ , \tag{4}$$

by definition, and we assume that

$$\frac{\partial \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\ln f(\mathbf{y};\boldsymbol{\theta})\right\}}{\partial\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\frac{\partial\ln f(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\} \tag{5}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The computation of the asymptotic p.d.f. is done using conditional probability density functions. The conditioning is on the event that the initial estimate is within the attraction region of the m'th maxima, which will be denoted by $\boldsymbol{\Theta}^m$, i.e.

$$f(\widetilde{\boldsymbol{\theta}}_N) = \sum_{m=0}^{M} f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)\mathbb{P}(\boldsymbol{\Theta}^m) \ , \tag{6}$$

where $f(\widetilde{\boldsymbol{\theta}}_N)$ is the distribution of $\widetilde{\boldsymbol{\theta}}_N$[1], $f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)$ is the distribution of $\widetilde{\boldsymbol{\theta}}_N$ given that the initial estimate was in $\boldsymbol{\Theta}^m$, and $\mathbb{P}(\boldsymbol{\Theta}^m)$ is the probability that the initial estimate was in $\boldsymbol{\Theta}^m$. The prior probabilities $\mathbb{P}(\boldsymbol{\Theta}^m)$ are assumed to be known in advance and can be found by empirical analysis of the initial estimator. These probabilities do not play a key role in the derivation or the applications discussed in the sequel. Here we implicitly assume that the entire space $\boldsymbol{\Theta}$ can be divided into disjoint subsets $\boldsymbol{\Theta}^m$, each of which is the attraction region of one of the maxima of $g(\boldsymbol{\theta}^0, \boldsymbol{\theta})$, and that $\bigcup_{m=0}^{M} \boldsymbol{\Theta}^m = \boldsymbol{\Theta}$.

For large N, given that the initial estimate is in $\boldsymbol{\Theta}^m$, $\widetilde{\boldsymbol{\theta}}_N$ is assumed to be in the close vicinity of $\boldsymbol{\theta}^m$, and the asymptotic conditional p.d.f. can be found using an analysis similar to that presented in [10] for the standard MLE and similar to Huber's derivation of the asymptotic p.d.f. of M-estimators [2]. The regularity conditions on $L_N(\mathbf{Y};\boldsymbol{\theta})$, which are needed for the derivation, are summarized in [3], and will be recalled during the derivation. One major difference of the present derivation from these other methods is that the Taylor expansion is performed around $\boldsymbol{\theta}^m$, which is not necessarily the true parameter, nor is it the global maximum (or minimum) of the target function. In order to give a self-contained treatment, we give the complete derivation for the case of a scalar parameter. For the case of a vector of parameters, we only state the final result.

3.1 Scalar Parameter Case

From the mean value theorem we have

$$\left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\widetilde{\theta}_N} = \left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\theta^m} + \left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} (\widetilde{\theta}_N - \theta^m) \ , \tag{7}$$

where $\theta^m < \overline{\theta} < \widetilde{\theta}_N$, assuming that the derivatives exist and are finite. Since $\widetilde{\theta}_N$ is a local maximum of the log-likelihood function, we have

$$\left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\widetilde{\theta}_N} = 0 \ . \tag{8}$$

Therefore,

$$\sqrt{N}(\widetilde{\theta}_N - \theta^m) = \frac{\frac{1}{\sqrt{N}} \left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\theta^m}}{-\frac{1}{N} \left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}}} \ . \tag{9}$$

[1] The dependency on the true parameter $\boldsymbol{\theta}^0$ has been omitted in order to simplify the notation.

Next, $\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}$ in the denominator is written explicitly

$$\frac{1}{N}\left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} = \frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log f(\mathbf{y}_n;\theta)}{\partial\theta^2}\right|_{\theta=\overline{\theta}} \quad . \tag{10}$$

Since $\theta^m < \overline{\theta} < \widetilde{\theta}_N$ and $\widetilde{\theta}_N \to \theta^m$ as $N \to \infty$ a.s., we must have $\overline{\theta} \to \theta^m$ as $N \to \infty$ a.s.. Hence

$$\begin{aligned}\frac{1}{N}\left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} &\to \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\left.\frac{\partial^2 \log f(\mathbf{y}_n;\theta)}{\partial\theta^2}\right|_{\theta=\theta^m}\right\} \quad \text{a.s.} \\ &\triangleq A(\theta^m) \ , \end{aligned} \tag{11}$$

due to the law of large numbers, where $\mathrm{E}_{\boldsymbol{\theta}^0}\left\{\left.\frac{\partial^2 \log f(\mathbf{y}_n;\theta)}{\partial\theta^2}\right|_{\theta=\theta^m}\right\}$ is assumed to be finite. In order to evaluate the numerator of (9), the following random variables are defined

$$x_n = \left.\frac{\partial \ln f(\mathbf{y}_n;\theta)}{\partial\theta}\right|_{\theta=\theta^m} \quad n = 1,\ldots,N \ . \tag{12}$$

Since the $\mathbf{y}_n$'s are independent and identically distributed, so are the x_n's. Therefore, by the Central Limit Theorem, the p.d.f. of the numerator of (9) will converge to a Gaussian p.d.f. with mean

$$\mathrm{E}_{\boldsymbol{\theta}^0}\left\{\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left.\frac{\partial \log f(\mathbf{y}_n;\theta)}{\partial\theta}\right|_{\theta=\theta^m}\right\} = 0 \tag{13}$$

and variance

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}^0}\left\{\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left.\frac{\partial \log f(\mathbf{y}_n;\theta)}{\partial\theta}\right|_{\theta=\theta^m}\right)^2\right\} &= \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\left(\left.\frac{\partial \log f(\mathbf{y}_n;\theta)}{\partial\theta}\right|_{\theta=\theta^m}\right)^2\right\} \\ &\triangleq B(\theta^m) \ , \end{aligned} \tag{14}$$

where we assume that $B(\theta^m)$ is finite. Next, Slutsky's theorem [11] is invoked. The theorem says that if x_n converges in distribution to x and z_n converges in probability to a constant c than x_n/z_n converges in distribution to x/c. Therefore, we arrive at the following result

$$\sqrt{N}(\widetilde{\theta}_N - \theta^m) \overset{\mathrm{a}}{\sim} N\left(0, \frac{B(\theta^m)}{A^2(\theta^m)}\right) \tag{15}$$

or, equivalently,

$$\widetilde{\theta}_N \overset{\mathrm{a}}{\sim} N\left(\theta^m, \frac{B(\theta^m)}{NA^2(\theta^m)}\right) \ , \tag{16}$$

where $\overset{\mathrm{a}}{\sim}$ denotes convergence in distribution. In the case where θ^m is the true parameter θ^0, we obtain the standard asymptotic Gaussian distribution of the MLE

$$\widetilde{\theta}_N \overset{\mathrm{a}}{\sim} N\left(\theta^o, I^{-1}(\theta^0)\right) \ , \tag{17}$$

where $I(\theta^0) = NA(\theta^0)$ is the Fisher Information (FI) of the measurements. However, it should be noted that in the general case $A(\theta^m) \neq -B(\theta^m)$.

In summary, the conditional p.d.f. $f(\widetilde{\theta}_N|\Theta^m)$ is asymptotically Gaussian with mean θ^m and variance $\frac{B(\theta^m)}{NA^2(\theta^m)}$, which equals $I^{-1}(\theta^0)$ only in the case where $m = 0$. Using this result, we can state that the asymptotic distribution of $\widetilde{\theta}_N$ in (6) is a Gaussian mixture with weights $\mathbb{P}(\Theta^m)$, $m = 0, \ldots, M$, which depend on the p.d.f. of the initial estimator.

3.2 Generalization to a Vector of Parameters

In the case of a vector of parameters, the conditional p.d.f. $f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)$ is asymptotically multivariate Gaussian with vector mean $\boldsymbol{\theta}^m$ and covariance matrix

$$\mathbf{C}_m \stackrel{\triangle}{=} \mathrm{Cov}_{\boldsymbol{\theta}^0}(\widetilde{\boldsymbol{\theta}}_N) = \frac{1}{N}\mathbf{A}^{-1}(\boldsymbol{\theta}^m)\mathbf{B}(\boldsymbol{\theta}^m)\mathbf{A}^{-1}(\boldsymbol{\theta}^m) \ , \tag{18}$$

which equals $\frac{1}{N}\mathbf{I}^{-1}(\boldsymbol{\theta}^0)$ - the Fisher Information Matrix (FIM) - in the case where $m = 0$, i.e. $\boldsymbol{\theta}^m$ is the global maximum. The kl elements of the matrices $\mathbf{A}(\boldsymbol{\theta})$ and $\mathbf{B}(\boldsymbol{\theta})$ are given by

$$\{\mathbf{A}(\boldsymbol{\theta})\}_{kl} = \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\frac{\partial^2 \log f(\mathbf{y}_n;\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l}\right\} \ , \tag{19}$$

and

$$\{\mathbf{B}(\boldsymbol{\theta})\}_{kl} = \mathrm{E}_{\boldsymbol{\theta}^0}\left\{\frac{\partial \log f(\mathbf{y}_n;\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial \log f(\mathbf{y}_n;\boldsymbol{\theta})}{\partial\theta_l}\right\} \ . \tag{20}$$

Therefore the asymptotic p.d.f. of $\widetilde{\boldsymbol{\theta}}_N$ is a multivariate Gaussian mixture.

The result (18) on the asymptotic conditional p.d.f. coincides with results reported in [4] in the context of misspecified models. Indeed, under the assumption $\widetilde{\boldsymbol{\theta}}_N \in \boldsymbol{\Theta}^m$, $m \neq 0$, the estimation problem can be viewed as a misspecified model. The family of distributions is correct but the domain of $\boldsymbol{\theta}$ does not contain the true parameter. In addition, the conditional p.d.f. $f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)$ can be found from Huber's work on M-estimators [2] by taking the target function that is minimized to be the negation of the log-likelihood function restricted to the attraction region of the specific local maximum.

The covariance (18) being equal to the inverse FIM is a necessary but not sufficient condition for $\boldsymbol{\theta}^m$ to be the global maximum. In particular, it is possible to construct a special parametric model in which $\mathbf{A}(\boldsymbol{\theta}^m)$ equals $-\mathbf{B}(\boldsymbol{\theta}^m)$ for $\boldsymbol{\theta}^m$ which is not the global maximum [5].

The following proposition summarizes the result presented in this section.

Proposition 1. *Under the assumptions made above, an estimator $\widetilde{\boldsymbol{\theta}}_N$ asymptotically follows a Gaussian mixture distribution with mean vectors $\boldsymbol{\theta}^m$ and covariance matrices $\mathbf{C}_m$ specified in* (18)*, i.e.*

$$f_{\widetilde{\boldsymbol{\theta}}_N}(\mathbf{t};\boldsymbol{\theta}^0) \to \sum_{m=0}^{M} \frac{\mathbb{P}(\Theta^m)}{(2\pi)^{K/2}\sqrt{|\mathbf{C}_m|}} \exp\left\{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\theta}^m)^T\mathbf{C}_m^{-1}(\mathbf{t}-\boldsymbol{\theta}^m)\right\}$$
$$\text{as } N \to \infty, \quad \forall\, \mathbf{t} \in \boldsymbol{\Theta} \ .$$

4 Applications

4.1 An Algorithm for Finding the MLE Based on the Asymptotic Distribution Result

In the present section, we propose an algorithm for finding the MLE that exploits the asymptotic results of the last section. As mentioned above, the algorithm was designed for scenarios in which the likelihood equations do not have a closed form solution, and, therefore, one must rely on iterative search over $\boldsymbol{\Theta}$ to find the MLE. If, in addition, the iterative search becomes computationally cumbersome for large data length, it might be impossible to perform the search algorithm on the log-likelihood function of the entire data set. In such cases, one can divide the complete data set into sub-blocks and find an estimator for each sub-block. These estimators will be referred to as sub-estimators. If the ambiguity function has one global maximum, then the average of the sub-estimators will closely approximate the MLE. However, if the ambiguity function has local maxima in addition to the global maximum, then some of the sub-estimators might converge to those local maxima and contribute large errors to the sub-estimators' average. A possible solution to this problem could be to cluster the sub-estimators and to choose the cluster whose members have the largest average log-likelihood value. However, if the dimension of the parameter vector is large and the local maxima of the ambiguity function are close to each other in $\boldsymbol{\Theta}$, the clustering problem becomes numerically intractable as well. Furthermore, as will be shown later, two remote local maxima might have nearly identical log-likelihood values. In such a case, the hight of the likelihood is not reliable for discriminating local from global maxima.

Therefore, we resort to a solution that circumvents the clustering requirement. To this end, we first employ the component-wise EM for mixtures (CEM) algorithm proposed by Figueiredo and Jain in [7]. Recall that according to the asymptotic result presented in the previous section, if the length of each data sub-block is large enough, the sub-estimators are random variables drawn from a Gaussian mixture distribution with means equal to the locations of the local maxima of the ambiguity function and covariance matrices as specified by (18). Therefore, the CEM can be used to estimate these mean and covariance parameters. The estimated means serve as candidates for the final estimate, and the estimated covariance matrices provide the means for discerning the global maximum using the procedure described below.

As can be seen from the derivation in Sec. 3, at the global maximum the covariance matrix of the estimates equals the inverse of the FIM. Therefore, in order to decide which local maxima are close to the global maximum, we can compare the estimated covariance matrices to the inverse of the FIM computed by an analytical or a numerical calculation, and choose the one having the best fit to this inverse FIM.

In order to explicitly state the algorithm, recall the statistical setting of our problem. The independent random vectors $\mathbf{y}_n$, $n = 1, \ldots, N$ have a common

p.d.f. $f(\mathbf{y};\boldsymbol{\theta})$, which is known up to the parameter vector $\boldsymbol{\theta}$ that is to be estimated. The algorithm is as follows:

1. Divide the entire data set into L sub-blocks of length N_{s}.
2. Find an estimator, which is a maximum of the log-likelihood of each of the sub-blocks, $\widehat{\boldsymbol{\theta}}^{l}_{N_{\mathrm{s}}}$; $l = 1, \ldots, L$, by some local optimization algorithm[2].
3. Run the CEM algorithm on $\widehat{\boldsymbol{\theta}}^{l}_{N_{\mathrm{s}}}$; $l = 1, \ldots, L$ to find the estimated means and covariance matrices of the Gaussian mixture model.
4. Compute either analytically or numerically the inverse of the FIM at each of the estimated means of the Gaussian mixture.
5. Choose the final estimate $\widehat{\boldsymbol{\theta}}_{final}$ to be the mean of the cluster that has the best fit between its estimated covariance and the inverse of the FIM evaluated at its mean (in the Forbenius norm sense, for example).

As for choosing the length N_{s} of the data sub-block, we will see in the simulations described below that the choice of N_{s} in the range of $\sqrt{N}$ gives the best results. Furthermore, since the covariance matrices of the clusters are known to be close to the inverse of the FIM, we use the FIM to initialize the CEM algorithm. Next, we present simulation results that validate the asymptotic p.d.f. stated in Prop. 1 and present a study of the performance of the proposed estimator.

4.2 Estimating Cauchy Parameters on a Non-Linear Manifold

Consider the following estimation problem, which is related to the estimation of a parameter, e.g. an image or a shape, embedded in a non-linear smooth manifold. The data are independent random vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ each of which is composed of three independent Cauchy random variables, with parameter $\alpha = 1$ and mode (median)

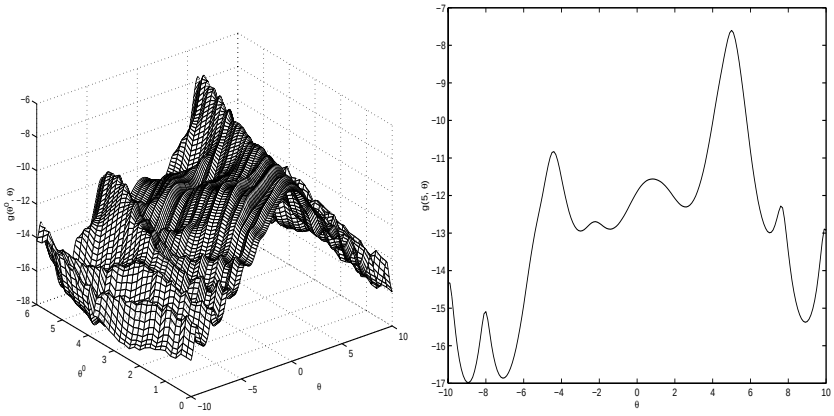
$$\boldsymbol{\mu}(\theta) = \begin{bmatrix} \mu_1(\theta) \\ \mu_2(\theta) \\ \mu_3(\theta) \end{bmatrix} = \begin{bmatrix} \theta \\ \theta \sin(\theta) \\ \theta \cos(\theta) \end{bmatrix} \quad , \tag{21}$$

i.e.,

$$f(y_i;\theta) = \frac{1/\pi}{1 + (y_i - \mu_i(\theta))^2}, \quad i = 1, 2, 3 \ . \tag{22}$$

These data can be considered as noisy measurements in $\mathbb{R}^3$ of the mode of the Cauchy density, which is constrained to lie on the manifold (a spiral) defined by (21). Since there exists no finite dimensional sufficient statistic for the mode of the Cauchy density, the complexity of the estimation problem increases in the number of samples. The ambiguity function associated with this estimation problem is depicted in Fig. 1(a) for different values of the true parameter θ^0, and a cross section is presented in Fig. 1(b) for $\theta^0 = 5$ - the value used in our simulations. Numerical calculations showed that the ambiguity function has two

[2] We assume that $\mathbb{P}(\boldsymbol{\Theta}^0) > 0$.



(a) The ambiguity function for different values of θ^0.

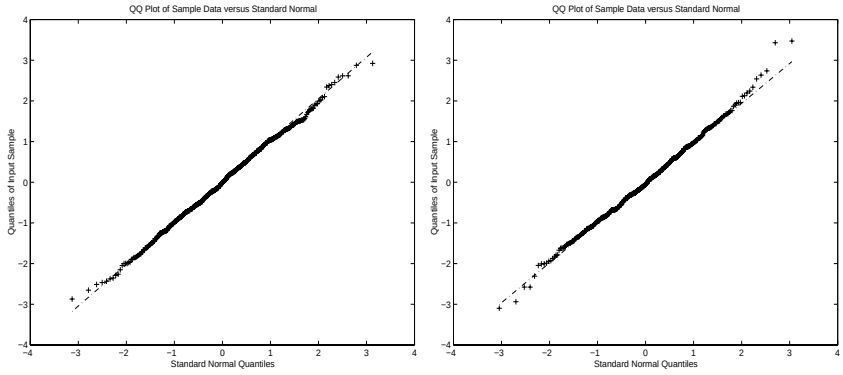
(b) Cross section of the ambiguity function at $\theta^0 = 5$.

Fig. 1. Multi-modal ambiguity function.

maxima in this region. One is the true parameter $\theta^0 = 5$ and another local maximum at $\theta^1 = 0.82$. Further analysis revealed that the regions of attraction associated with these modes are the open intervals $\Theta^0 = (2.56, 6)$ and $\Theta^1 = (0, 2.56)$, respectively. In addition, the analytical result (16) predicts that in cases where the search algorithm converges to θ^0, the estimate will be Gaussian with mean θ^0 and variance $\frac{B(\theta^0)}{NA^2(\theta^0)} = \frac{1}{NA(\theta^0)} = \frac{0.074}{N}$, and in cases where the search algorithm converges to θ^1, the estimate will be Gaussian with mean θ^1 and variance $\frac{B(\theta^1)}{NA^2(\theta^1)} = \frac{0.31}{N}$. Since the initial estimate is uniformly distributed, it is easily found that $\mathbb{P}(\Theta^0) = 0.57$ and $\mathbb{P}(\Theta^1) = 0.43$. In practice, these values are estimated by the CEM algorithm, even though they play no role in determining the final estimate.

In our simulations, $N = 200$ and the local optimization algorithm is Matlab's routine 'fminsearch', which implements the Nelder-Mead algorithm on the log-likelihood function. The starting point for the algorithm is chosen randomly in the interval $[0, 6]$. 1000 Monte Carlo trials showed good agreement with the analytical predictions (16). In order to verify the Gaussian mixture distribution of the estimates, they were divided into two groups, one contained the estimates that were around θ^0 and the second contained the estimates around θ^1. Then, the two groups were centralized according to the predicted mean, divided by the predicted standard deviation, and compared against the standard Gaussian distribution. The resulting Q-Q plots are depicted in Figs. 2(a) and 2(b).

Next, the performance of this algorithm was examined. The entire data record was divided into sub-blocks for several choices of block lengths. The CEM was used to find the estimated number of clusters, their means, and variances. The



(a) Estimates around $\theta^0 = 5$ normalized according to $\frac{B(\theta^0)}{NA^2(\theta^0)} = \frac{1}{NA(\theta^0)}$.

(b) Estimates around $\theta^1 = 0.82$ normalized according to $\frac{B(\theta^1)}{NA^2(\theta^1)}$.

Fig. 2. Validation of the Gaussian mixture distribution.

variance of each cluster was compared to the inverse of the Fisher information at the mean of each cluster. The Fisher information for this statistical problem can be found analytically to be $I(\theta) = \frac{2+\theta^2}{2\alpha^2}$. The final estimate was the mean of the cluster that its variance was closer to the inverse of $I(\theta)$ evaluated at the mean.

The probability of deciding on the wrong maximum, which will be referred to as the probability of large error, and the small error performance in cases where the decision was correct were estimated using 500 Monte Carlo trials. As expected, the small error performance improved as the number of samples in each sub-block increases. However, the probability of a large scale error has a minimum point with respect to the sub-block length as seen in Fig. 3. Thus, there is an optimum sub-block length for minimizing the influence of large errors. An intuitive explanation of this phenomenon is the following. When the sub-block size is too large, the Gaussian mixture approximation is good but the number of samples available for the CEM estimation is small, resulting in poor covariance estimation which leads to estimation errors. On the other hand, when the number of sub-blocks is large the amount of data available to the CEM algorithm is large. However, since the number of samples at each sub-block is small, the data are far from being distributed as a Gaussian mixture, and the variance of the estimator around the true parameter no longer equals the inverse of the Fisher information, which again results in estimation errors.

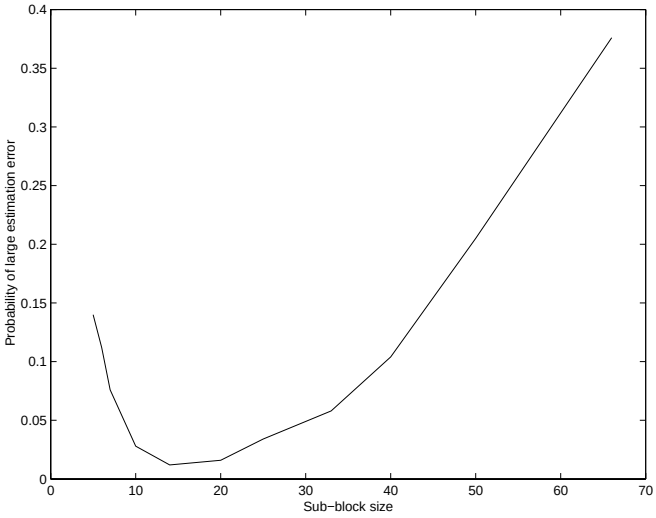


Fig. 3. Influence of sub-block length on the probability of large error. Note that the optimal block length is $\sqrt{N}$, where here $N = 200$.

4.3 Aggregation of Estimates from Remote Sensors

The present section, addresses a scenario in which the division of the entire data sample into sub-blocks is imposed by the system design. Consider the following distributed processing problem. A large number of low power sensors are geographically distributed in order to perform an estimation task. Each of these sensors collects data generated independently by a common parametric model $f(\mathbf{y};\boldsymbol{\theta})$, which has a multi-modal ambiguity function. Due to power and bandwidth constraints, the sensors do not transmit the complete data to the central processing unit, but rather each performs a suboptimal local search on the log-likelihood function and transmits only its local estimate. The question then arises as to how to treat the large number of estimates, some of which may correspond to successful convergence to the global maximum and some to erroneous local maxima.

Again, the analytical result stated in Prop. 1 provides the means to find a global estimate through a well-posed Gaussian mixture problem. The data available for the central processing unit are the local estimates delivered by the individual sensors. The theory in Sec. 3 asserts that these are drawn from a Gaussian Mixture model. Furthermore, the cluster corresponding to estimates which are close to the global maximum has the property that its covariance matrix is close to the inverse of the FIM evaluated at mean of this cluster of estimates. As in the previous application, this property will be used to find the final estimate.

Simulation Results. We generate 2D sensors data from the following Gaussian mixture density

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{2} \alpha_j f(\mathbf{y};\boldsymbol{\mu}_j) \ , \tag{23}$$

where $f(\mathbf{y};\boldsymbol{\mu}_j)$ is the bivariate Gaussian density

$$f(\mathbf{y};\boldsymbol{\mu}_j) = \frac{1}{2\pi\sqrt{|\mathbf{C}_j|}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_j)^T \mathbf{C}_j^{-1}(\mathbf{y}-\boldsymbol{\mu}_j)\right\} \ , \tag{24}$$

and where $\mathbf{y} = [y_1 \ y_2]^T$. Note that the Gaussian mixture model of the new data (23) has nothing to do with the Gaussian mixture model which is an asymptotic distribution for the local estimates $\widehat{\boldsymbol{\theta}}_N^l$; $l = 1, \ldots, L$. The parameters vector $\boldsymbol{\theta}$ contains the two vector means $\boldsymbol{\mu}_j = [\mu_{j1} \ \mu_{j2}]^T$; $j = 1, 2$ in the following order

$$\boldsymbol{\theta} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \ . \tag{25}$$

The entries of the covariance matrices $\mathbf{C}_j$; $j = 1, 2$ associated with each of the components and the mixing probabilities α_1 and α_2 are assumed known. This is a simple model corresponding to a network of L 2D position estimating sensors.

Each sensor estimates $\boldsymbol{\theta}$ from $N = 50$ samples. The true values for the location parameters to be estimated were chosen to be

$$\boldsymbol{\theta}^0 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{bmatrix} \ . \tag{26}$$

The remaining known parameters were chosen to be

$$\mathbf{C}_1 = \mathbf{C}_2 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix} \tag{27}$$

and

$$\alpha_1 = 0.4; \qquad \alpha_2 = 0.6 \ . \tag{28}$$

The vector means are known a-priori to lie in the rectangle $\boldsymbol{\Theta} = \{[0,3] \times [0,3]\}$. Typical sensor data, generated according to the above model (23) are presented in Fig. 4. The two circles correspond to the two components.

Each sensor uses the following algorithm to find an estimate. A point is generated randomly, according to a uniform distribution on the given rectangle $\boldsymbol{\Theta}$. Then this point is used as the starting point of a local search for a maximum of the log-likelihood function of the measurement. In our simulation, we used the Matlab routine 'fminsearch' which applies the Nelder-Mead algorithm to

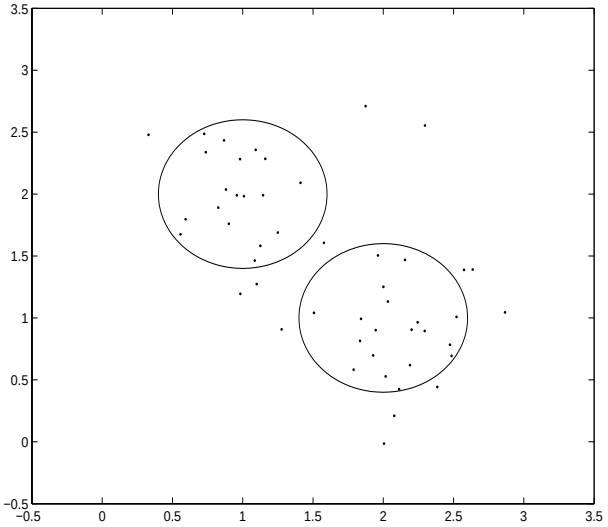


Fig. 4. Measured data for a single sensor.

maximize the local log-likelihood function $L_N(\mathbf{Y};\boldsymbol{\theta})$ with respect to the unknown parameters $\boldsymbol{\theta}$. Denote the estimate from the l'th sensor by $\widehat{\boldsymbol{\theta}}_N^l$.

We have found that the ambiguity function has two maxima in $\boldsymbol{\Theta}$. One maximum is at the true parameters vector $\boldsymbol{\theta}^0$ and the second maximum

$$\boldsymbol{\theta}^1 = \begin{bmatrix} 2.05 \\ 0.95 \\ 1.08 \\ 1.92 \end{bmatrix} \tag{29}$$

corresponds to the reversed model, i.e., switching between the two components. Therefore, the estimates $\widehat{\boldsymbol{\theta}}_N^l;\ l = 1,\ldots,L$ available at the processing unit can be seen as samples drawn from a two component multi-variate (4-dimensional) Gaussian mixture, where the vector means of the two components are the locations of the two maxima of the ambiguity function in the parameters space and the covariance matrices are as presented in (18). The 4-dimensional estimates generated by $L = 200$ sensors are presented in the Figs. 5(a) and 5(b). Each sub-figure corresponds to two parameters. In each figure, the circled cluster correspond to estimates that are close to the global maximum and the second cluster corresponds to estimates that are close to the local maximum.

An intuitive approach for clustering the two groups of estimates could be to use the actual values of the log-likelihood at the point of convergence, which could be transmitted in addition to the estimates to the central processing unit. However, since the mixing probabilities $\{\alpha_1,\ \alpha_2\}$ are close to $\{1/2,\ 1/2\}$, the two components are similar and the value of the log-likelihood function at the global and local maxima are nearly identical. This phenomenon renders impossible

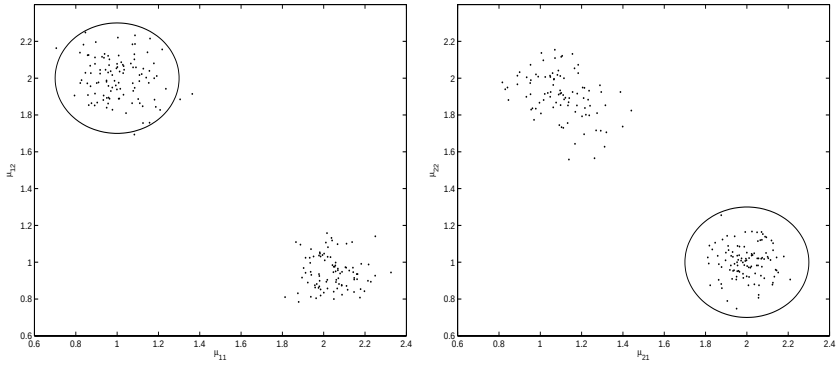


Fig. 5. Estimates $\widehat{\boldsymbol{\theta}}_N^l$; $l = 1, \ldots, L$ generated by $L = 200$ sensors.

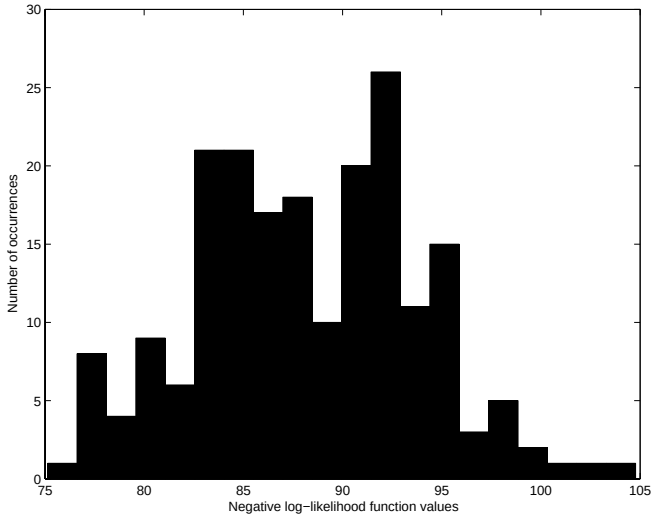


Fig. 6. Histogram of the log-likelihood function values $\ln f(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_N^l)$; $l = 1, \ldots, L$ obtained from estimates $\widehat{\boldsymbol{\theta}}_N^l$; $l = 1, \ldots, L$ generated by $L = 200$ sensors.

the discrimination between 'good' estimates (global maximum) and erroneous ones (local maximum), using only the log-likelihood function values. In Fig. 6 a histogram of the negative log-likelihood function values $\ln f(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_N^l)$; $l = 1, \ldots, L$ from one simulation is presented. It is not clear from this histogram that there are two separable components.

In contrast, we can reliably discriminate between the two local maxima based on the curvature of the parametric model at each local maxima. As was shown in Sec. 3, the covariance matrices of the two components of estimates are directly

related to the curvature of the ambiguity function at the two maxima, and at the global maximum equal the inverse of the FIM. Therefore, the algorithm proposed in Sec. 4.1 can be applied.

First the number of components, the mean vectors and the covariance matrices, of the estimates are estimated using the CEM algorithm. The estimated mean vectors serve as candidates for the final estimate and the estimated covariance matrices provides the means to find the component that corresponds to the global maximum. More explicitly, for each component the distance between the estimated covariance and the inverse of the FIM calculated at the point of the mean is computed. In our simulation, the Frobenius norm of the difference matrix was used as the distance measure. Finally, the mean of the component with the smallest norm is chosen as the final estimate. Since the 4×4 dimensional FIM cannot be computed analytically, it is computed by numerical integration and then inverted. The kl entry of the FIM is found by numerically calculating the following integral

$$\mathrm{FIM}_{kl} = \int_{-\infty}^{\infty} \frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \frac{\partial \log f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_l} f(\mathbf{y};\boldsymbol{\theta}) \mathrm{d}\mathbf{y} \ , \tag{30}$$

where the estimated mean of the candidate component is plugged-in for the unknown parameters.

The algorithm was tested in the above setting for several possible numbers of sensors L in order to evaluate two aspects of its performance. The first is the probability of detecting the global maximum. The second is the small-scale estimation errors when the global maximum is detected correctly. The algorithm was run 100 times for $L = 50, 100, 150$ and 200 sensors. In the case of $L = 50$ sensors, there were 6 cases of erroneous decisions. For systems of $100, 150$, and 200 sensors there was 100 percent success, i.e. the algorithm always detected the correct maximum. The fact that the estimated covariance matrix of the two components is small, which is usually the case when the number of samples at each sensor is sufficiently large, contributed to the success of the CEM stage. The small-scale estimation errors in cases where the global maximum was detected correctly, were compared to the performance of a clairvoyant estimator which knows which local estimates are close to the global max. This clairvoyant estimator averages only those estimates that close to the global maximum. The performances of the CEM estimator and the clairvoyant estimator are identical.

5 Concluding Remarks

The work presented in this paper is closely related to the work of White [4] on misspecified models and to the work of Gan and Jiang [5] on the problem of local maxima. Given a ML estimate, White proposed a test to detect a misspecified model. Given a local maximizer of the log-likelihood function, Gan and Jiang offered the same test in order to detect a scenario of convergence to a local maximum. This test is based on the observation that the two ways to estimate the FIM from the data given the estimated parameters, i.e., the Hessian form and the

outer product form, converge to the same value in the case of a global maximum in a correctly specified model. The test statistic, which is the difference between those two estimators of the FIM, was shown to be asymptotically Gaussian distributed. However, as mentioned in [5], the convergence of the test statistic to its asymptotic distribution is slow and the test suffers from over rejection in a moderate number of samples. Therefore, this test could not be used to determine whether or not the sub-estimates of the algorithm proposed in Sec. 4.1 are related to a global maximum. Furthermore, this test requires access to the data and therefore, could not be used in the estimates fusion problem, discussed in Sec. 4.3.

In contrast, the present paper considers cases in which the complete data are divided into sub-blocks, either due to computational burden or due to the system design. This data partitioning gives direct access to the estimated covariance matrix of the sub-estimates, which can then be compared to the calculated FIM. This procedure has considerably better performance and does not require re-processing the complete data.

6 Acknowledgement

This work was partially supported by a Dept. of EECS at the University of Michigan Fellowship and DARPA-MURI Grant Number DAAD19-02-1-0262.

References

1. R. A. Fisher. On the mathematical foundation of theoretical statistics. *Phil. Trans. Roy. Soc. London*, 222:309–368, 1922.
2. P.J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.
3. P.J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, 1967.
4. H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–26, Jan 1982.
5. L. Gan and J. Jiang. A test for global maximum. *Journal of the American Statistical Association*, 94(447):847–854, Sep 1999.
6. A.K. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–38, Jan 2000.
7. M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixute models. *IEEE Trans on Pattern Anal and Machine Intelligence*, 24:381–396, March 2002.
8. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data using the em algorithm. *Ann. Roy. Statist. Soc.*, 39:1–38, Dec 1977.
9. A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 60:595–603, Dec 1949.
10. S.M. Kay. *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall, 1993.
11. P.J. Bickel and K.A. Doksum. *Mathematical Statistics*. Holden-Day, San Francisco, 1977.