Asymptotic Distribution of Log-Likelihood Maximization Based Algorithms and Applications

D. Blatt, A. Hero

Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, MI
e-mail: 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 iterative search, are analyzed. 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. In addition, an algorithm for finding the maximum likelihood estimator is offered. 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 result on the asymptotic distribution is validated and the performance of the offered algorithm is examined by computer simulations.

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 Normal 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 an estimate 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.

In the first part of this paper, the asymptotic distribution of estimates that are based on a sub-optimal search for the ML estimate 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 iterative search that converges to a maximum point, are analyzed and shown to be asymptotically mixed Normal distributed. The result is linked to previous results by Huber [3] and White [4] as explained in detail below.

In the second part of the paper, an algorithm based on this result is offered. 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. The second stage involves the estimation of a finite mixture model, which is a classic problem in statistical pattern recognition (e.g. [5], [6], and references therein).

2 Problem Formulation

The independent random vectors $\mathbf{y}_n$, $n = 1, \ldots, N$ have a common probability density function (pdf) $f(\mathbf{y}, \boldsymbol{\theta})$, 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 analytic 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 [7] and its variations. This method will be referred to as the two-stage method, 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 pdf 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 $\boldsymbol{\Theta}$.

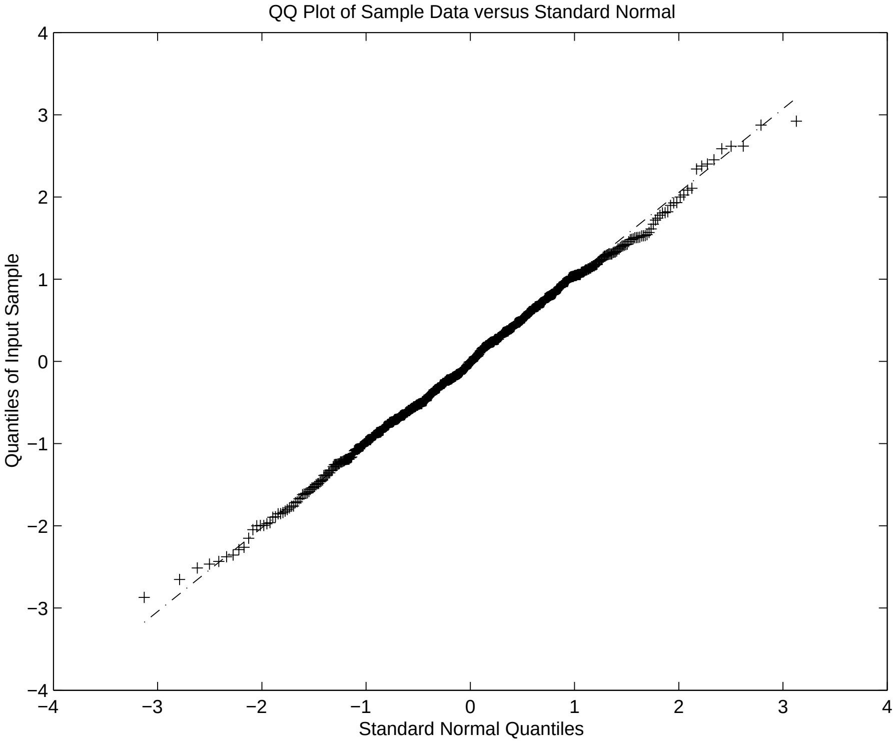

Fig. 1. Estimates around $\theta^0 = 5$ normalized according to $\frac{B(\theta^0)}{NA^2(\theta^0)} = \frac{1}{NA(\theta^0)}$

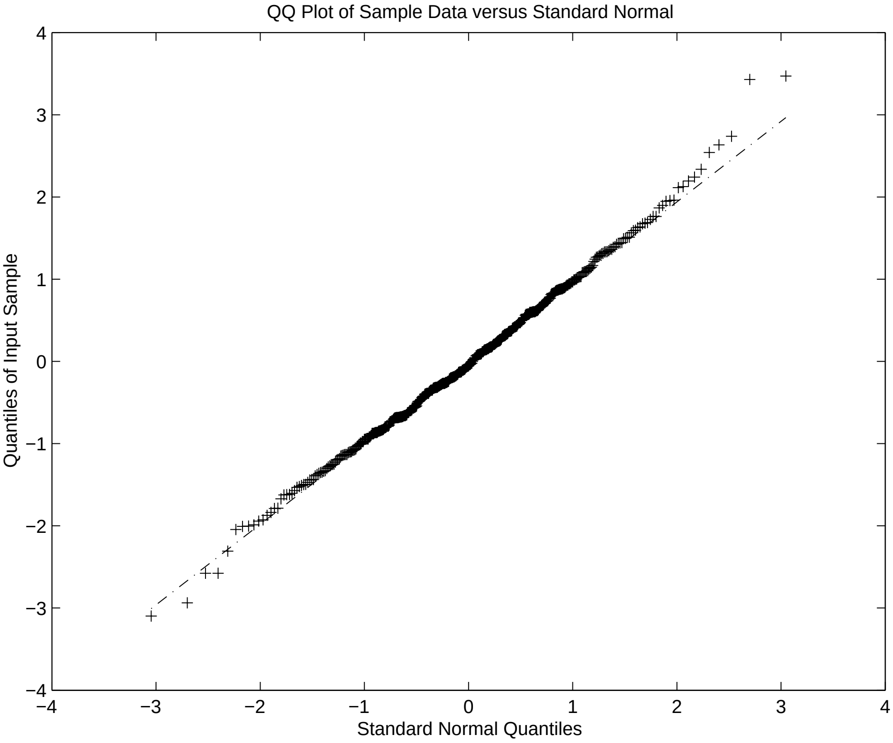

Fig. 2. Estimates around $\theta^1 = 0.82$ normalized according to $\frac{B(\theta^1)}{NA^2(\theta^1)}$

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

$$g(\boldsymbol{\theta}^0,\boldsymbol{\theta}) = \mathrm{E}\left\{\ln f(\mathbf{y};\boldsymbol{\theta})\right\} \overset{\triangle}{=} \int_{\mathcal{Y}} \ln\left(f(\mathbf{y};\boldsymbol{\theta})\right) f(\mathbf{y};\boldsymbol{\theta}^0) d\mathbf{y}, \tag{3}$$

where $\mathrm{E}\left\{\cdot\right\}$ denotes the statistical expectation with respect to the true parameter $\boldsymbol{\theta}^0$. 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 $\boldsymbol{\theta}^0$ [8], and it is assumed to have a number of local maxima at points which will be noted by $\boldsymbol{\theta}^m$, $m = 1,\ldots,M$. All the local maxima satisfy

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

by definition.

The computation of the asymptotic pdf is done using conditional pdfs. 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.

$$f(\widetilde{\boldsymbol{\theta}}_N) = \sum_{m=0}^{M} f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)\mathbb{P}(\boldsymbol{\Theta}^m), \tag{5}$$

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 in the performance analysis of the initial estimate. Here we implicitly assume that the entire space $\boldsymbol{\Theta}$ can be divided into disjoint subsets, 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 pdf can be found using an analysis similar to that presented in [9] for the standard MLE and to Huber's derivation of the asymptotic pdf of the M-estimators [2]. The regularity conditions on $L_N(\mathbf{Y};\boldsymbol{\theta})$, which are needed for the derivation, are summarized in [3]. One major difference of the present derivation from these methods is that the Taylor expansion is performed around $\boldsymbol{\theta}^m$, which is not necessarily the true parameter. In order to give a self-contained treatment, we give the

[1] The dependency on the true parameter has been omitter in order to simplify the notation.

complete derivation for the case of a scalar parameter. For the case of a vector of parameters, we state the final result.

From the mean value theorem we have

$$\left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\widetilde{\theta}_N} = \left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\theta^m} + \left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} (\widetilde{\theta}_N - \theta^m) \tag{6}$$

where $\theta^m < \overline{\theta} < \widetilde{\theta}_N$. Since $\widetilde{\theta}_N$ is a maximum point of the log-likelihood function, we have

$$\left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\widetilde{\theta}_N} = 0. \tag{7}$$

Therefore, we get

$$\sqrt{N}(\widetilde{\theta}_N - \theta) = \frac{\frac{1}{\sqrt{N}}\left.\frac{\partial L_N(\mathbf{Y};\theta)}{\partial\theta}\right|_{\theta=\theta^m}}{-\frac{1}{N}\left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}}}. \tag{8}$$

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

$$\frac{1}{N}\left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} = \frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log f(\mathbf{y}_n,\theta)}{\partial\theta^2}\right|_{\theta=\overline{\theta}}. \tag{9}$$

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

$$\begin{aligned}\frac{1}{N}\left.\frac{\partial^2 L_N(\mathbf{Y};\theta)}{\partial^2\theta}\right|_{\theta=\overline{\theta}} &\to \frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log f(\mathbf{y}_n,\theta)}{\partial\theta^2}\right|_{\theta=\theta^m} \\ &\to \mathrm{E}\left\{\left.\frac{\partial^2 \log f(\mathbf{y}_n,\theta)}{\partial\theta^2}\right|_{\theta=\theta^m}\right\} \quad a.s. \\ &\triangleq A(\theta^m), \end{aligned} \tag{10}$$

where the last convergence is due to the law of large numbers. In order to evaluate the numerator, the following random variables are defined

$$x_n = \left.\frac{\partial \ln f(\mathbf{y}_n,\theta)}{\partial\theta}\right|_{\theta=\theta^m} \quad n = 1,\ldots,N. \tag{11}$$

Since the $\mathbf{y}_n$'s are independent and identically distributed, so are the x_n's. Therefore, by the Central Limit Theorem , the pdf of the numerator in (8) will converge to a Normal pdf with mean

$$\mathrm{E}\left\{\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left.\frac{\partial \log f(\mathbf{y}_n,\theta)}{\partial\theta}\right|_{\theta=\theta^m}\right\} = 0 \tag{12}$$

and variance

$$\mathrm{E}\left\{\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\frac{\partial\log f(\mathbf{y}_n,\theta)}{\partial\theta}\bigg|_{\theta=\theta^m}\right)^2\right\}=\mathrm{E}\left\{\left(\frac{\partial\log f(\mathbf{y}_n,\theta)}{\partial\theta}\bigg|_{\theta=\theta^m}\right)^2\right\}$$
$$\triangleq B(\theta^m). \tag{13}$$

Next, using Slutsky's theorem [10], we arrive at the following result

$$\sqrt{N}(\widetilde{\theta}_N-\theta^m)\overset{a}{\sim}N\left(0,\frac{B(\theta^m)}{A^2(\theta^m)}\right) \tag{14}$$

or, equivalently,

$$\widetilde{\theta}_N\overset{a}{\sim}N\left(\theta^m,\frac{B(\theta^m)}{NA^2(\theta^m)}\right). \tag{15}$$

In the case where θ^m is the true parameter θ^0, we get the standard asymptotic normality of the MLE

$$\widetilde{\theta}_N\overset{a}{\sim}N\left(\theta^o,I^{-1}(\theta^0)\right), \tag{16}$$

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 pdf $f(\widetilde{\theta}_N|\Theta^m)$ is asymptotically Normal 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 (5) is a Normal mixture with weights $\mathbb{P}(\Theta^m)$, $m=0,\ldots,M$, which depend on the initial estimator's performance.

3.1 Generalization for a Vector of Parameters

In the case of a vector of parameters, the conditional pdf $f(\widetilde{\boldsymbol{\theta}}_N|\boldsymbol{\Theta}^m)$ is asymptotically multivariate Normal with vector mean $\boldsymbol{\theta}^m$ and variance

$$\frac{1}{N}\mathbf{A}^{-1}(\boldsymbol{\theta}^m)\mathbf{B}(\boldsymbol{\theta}^m)\mathbf{A}^{-1}(\boldsymbol{\theta}^m), \tag{17}$$

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

$$\mathbf{A}(\boldsymbol{\theta})=\left\{\mathrm{E}\left\{\frac{\partial^2\log f(\mathbf{y}_n,\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right\}\right\}, \tag{18}$$

and

$$\mathbf{B}(\boldsymbol{\theta})=\left\{\mathrm{E}\left\{\frac{\partial\log f(\mathbf{y}_n,\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial\log f(\mathbf{y}_n,\boldsymbol{\theta})}{\partial\theta_l}\right\}\right\}. \tag{19}$$

Therefore the asymptotic pdf of $\widetilde{\boldsymbol{\theta}}_N$ is a multivariate Normal mixture.

The result on the asymptotic conditional pdf 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 pdf 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 likelihood function restricted to the attraction region of the specific local maximum.

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

In the present section we offer an algorithm for finding the MLE. As mentioned above, the algorithm was designed for scenarios in which the likelihood equations do not have a closed form solution, and, therefore, one seeking the MLE must rely on a search algorithm. If, in addition, the iterative search is computationally cumbersome and highly dependent on the data length, it might be impossible to find the MLE for the entire data set. In such cases, one can divide the complete data set into sub-blocks and find the MLE for each sub-block. These estimators will be referred to as sub-MLEs. If the ambiguity function has one global maximum, then the average of the sub-MLEs will make a good solution to the complete problem. 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-MLEs' average. A possible solution to this problem would be to cluster the sub-MLEs and to choose the cluster whose members have the largest average 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, the clustering problem becomes numerically intractable as well.

Therefore, we resort to a solution that circumvents the clustering requirement. To this end, we first employ the the component-wise EM for mixtures (CEM^2) algorithm offered by Figueiredo and Jain in [6]. Recall that according to the result presented in the previous section, if the length of each data sub-block is large enough, the sub-MLEs are random variables drawn from a Normal mixture distribution with means equal to the local maxima of the ambiguity function and covariance matrices as specified in 3.1. Therefore, the model of the mixture is known up to parameters that can be estimated directly using CEM^2 without the need for an actual clustering of the sub-MLEs. The estimated means serve as candidates for the final estimate, and the estimated covariances provide the means for choosing the correct mean using the following property.

As observed first by White [4] in his work on misspecified models, later by Gan and Jiang [11] specifically for the problem of local maxima, and can also be seen from the derivation in section 3, only at the global maximum does the expected value of the Hessian of the likelihood function equal the negation of the expected value of the outer product of the first derivatives and, therefore, the covariance matrix of the estimates equals the inverse of the FIM. Therefore, in order to decide which local maximum is the global maximum, we can compare the estimated covariance matrices with the analytical calculation of the inverse of the FIM and choose the one with best proximity.

It should be noted that the tests offered in [4] and [11] cannot be directly applied using the estimated means, since these may not necessarily be local maxima of the complete data set, and, therefore, the assumption that the test is performed on a local maximum of the complete data is violated. In fact, as seen from our simulations, this violation tampers the validity of the test and leads to over-rejection. In other words, only if several maximizations of the likelihood of the entire data are possible, the estimated means of the mixture model can be viewed as intelligent starting points, which should be tested one by one until the test for a global maximum [11] is passed.

In order to explicitly state the algorithm, consider once more the following statistical problem. The independent random vectors $\mathbf{y}_n$, $n = 1, \dots, N$ have a common probability density function (pdf) $f(\mathbf{y}, \boldsymbol{\theta})$, which is known up to the parameter vector $\boldsymbol{\theta}$ that is to be estimated. The algorithm is stated as follows:

1. Divide the entire data set into sub-blocks of length L.
2. Find the MLE of each of the sub-blocks $\widehat{\boldsymbol{\theta}}_l$, $\quad l = 1, \dots, L$, by using any initial rough estimator[2].
3. Run the CEM^2 algorithm on $\{\widehat{\boldsymbol{\theta}}_l\}_{l=1}^{L}$ and find the estimated means and covariance matrices of the Normal mixture model.
4. Compute the inverse of the FIM at each of the estaimated means.
5. Choose the final estimate $\widehat{\boldsymbol{\theta}}_{final}$ to be the vector that minimizes the distance between the estimated and the calculated covariance matrices (in the Forbenius norm sense, for example).

As for choosing the length of the data sub-block, we have seen in our simulations that the choice of L in the range of $\sqrt{N}$ gives the best results. However, this still remains an open research question. Furthermore, since the covariance matrices are known to be at least close to the inverse of the FIM, we can use this information in the initialization of the CEM^2 algorithm. In the following section, we present simulation results that validate the asymptotic pdf derived in section 3 and present a performance analysis for the offered algorithm for a specific estimation problem.

5 Simulation Results

Consider the following estimation problem, which can be seen as a simplification of problems related to the embedding of images in a manifold. The data are independent random vectors $\mathbf{y}_1$, $\mathbf{y}_2$, $\dots$, $\mathbf{y}_N$ each of which is composed of three independent Cauchy random variables, with parameter $\alpha = 1$ and mode equals

$$\boldsymbol{\mu}(\theta) = \begin{bmatrix} \mu_1(\theta) \\ \mu_2(\theta) \\ \mu_3(\theta) \end{bmatrix} = \begin{bmatrix} \theta \\ \theta \sin(\theta) \\ \theta \cos(\theta) \end{bmatrix}, \tag{20}$$

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

i.e.,

$$f(y_i;\theta) = \frac{1/\pi}{(y_i - \mu_i(\theta))^2 - 1}, \quad i = 1, 2, 3. \tag{21}$$

These data can be considered as measurements in $\mathbb{R}^3$ that have a mode, which is on the manifold (a spiral) defined by (20). Since there exists no finite dimensional sufficient statistic, the complexity of the estimation problem is highly dependent on the number of samples. The ambiguity function associated with this estimation problem is depicted in figure 3 for different values of the true parameter θ^0, and a cross section is presented in figure 4 for $\theta^0 = 5$ - the value used in our simulations. The initial estimator is assumed to produce an estimate which is uniformly distributed over $[0, 6]$. Numerical calculations showed that the ambiguity function has two maxima in this region. One is the true parameter $\theta^0 = 5$ and another maximum at $\theta^1 = 0.82$. Further analysis revealed that $\Theta^0 = (2.56, 6)$ and $\Theta^1 = (0, 2.56)$. In addition, the analytical result predicts that in cases where the search algorithm converges to θ^0, the estimate will be Normal 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 Normal 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 our simulations, $N = 200$ and the initial estimate is used as the starting point for Matlab's routine 'fminsearch', which implements the Nelder-Mead algorithm. 1000 Monte Carlo trials showed good agreement with the analytical results. The estimates were divided into two groups, one contained the estimates that were around θ^0 and the other group 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 Normal distribution. The resulting Q-Q plots are depicted in figures 1 and 2.

Next, the performance of the offered algorithm was examined. The entire data record was divided into sub-blocks for several choices of block lengths. The CEM^2 was used to find the estimated number of clusters their means and variances. The variance of each cluster was compared to the inverse of the FI at the mean of each cluster. The FI 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 $I(\theta)$.

The probability of deciding on the wrong maximum 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 in cases of estimating the correct maximum 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 figure 5. An intuitive explanation of this phenomenon can be the following. When the sub-block size is too large, the Normal mixture approximation is good but the number of samples available for the CEM^2 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^2 algorithm is large. However, since the number of samples at each sub-block is small, the data are far from being distributed as Normal mixture, and the variance of the estimator around the true parameter no longer equals the inverse of the FI, which again results in estimation errors.

6 Concluding Remarks

In the present paper, we derived the asymptotic distribution of estimators that are based on an iterative maximization of the likelihood function that may converge to a local maximum. Based on this result, we designed an algorithm for finding the MLE with good reliability in scenarios where the iterative computation of the MLE is computationally cumbersome. Finally, simulation results validated the analytical results and characterized the performance of the offered algorithm.

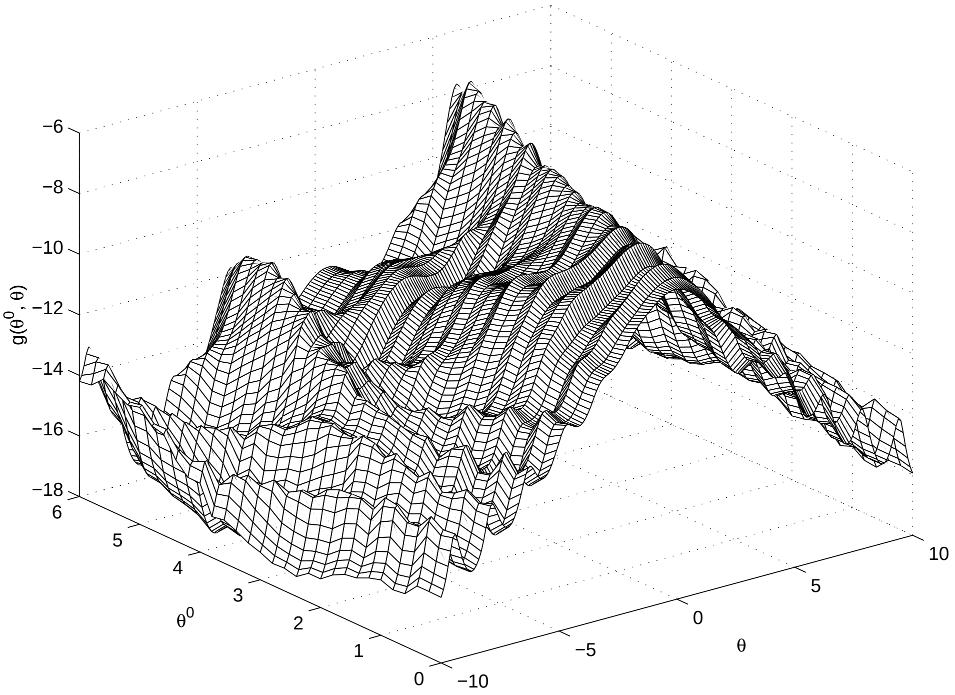

Fig. 3. The ambiguity function

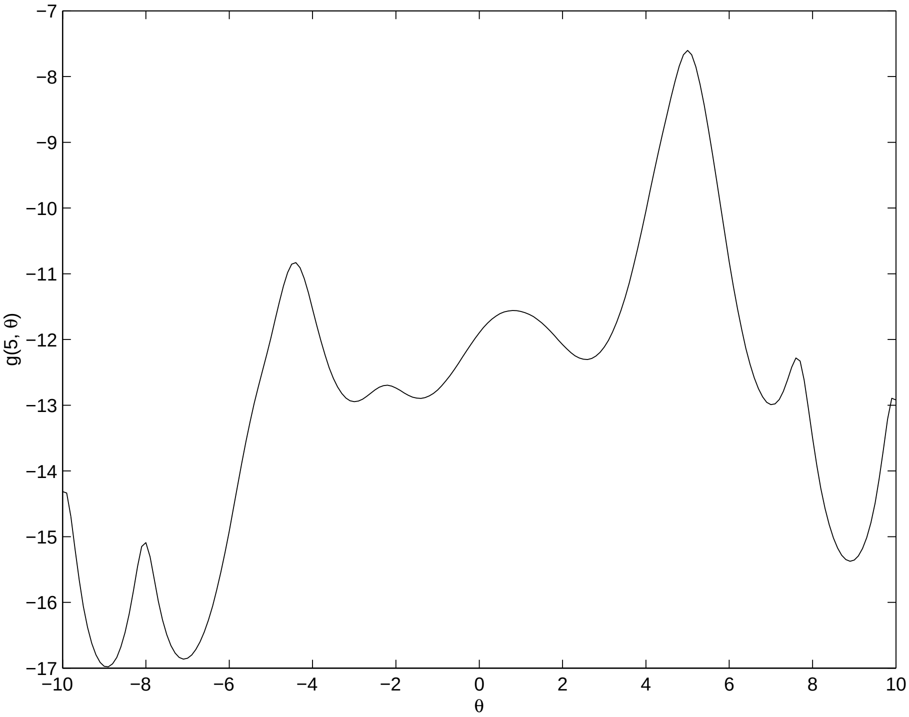

Fig. 4. Cross section of the ambiguity function at $\theta^0 = 5$

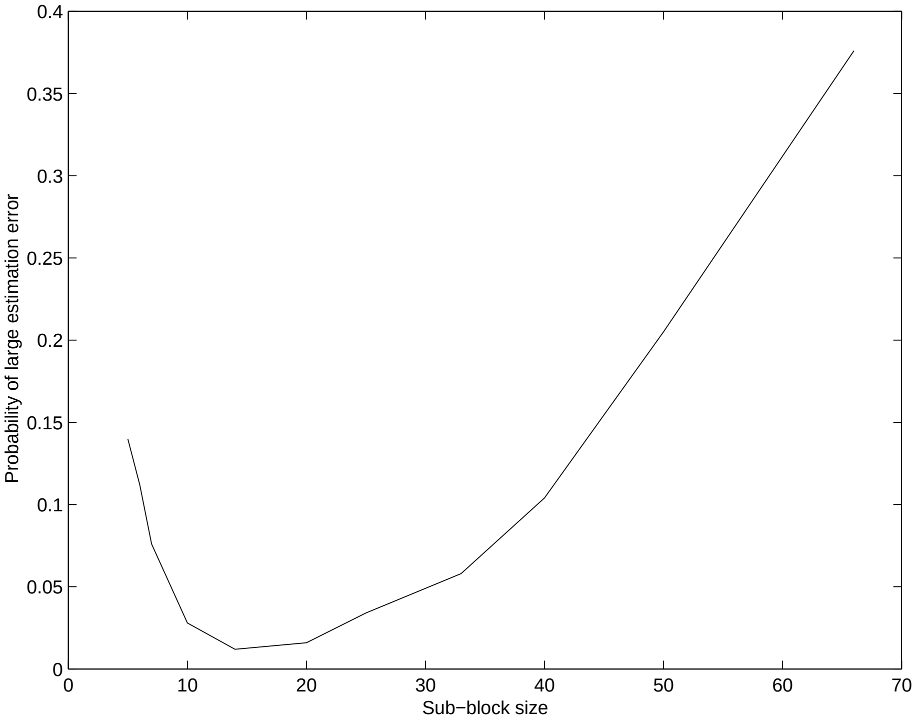

Fig. 5. Probability of large estimation error for several sub-block lengths

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. 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.
6. 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.
7. 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.
8. A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 60:595–603, Dec 1949.
9. S.M. Kay. *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall, 1993.
10. P.J. Bickel and K.A. Doksum. *Mathematical Statistics*. Holden-Day, San Francisco, 1977.
11. L. Gan and J. Jiang. A test for global maximum. *Journal of the American Statistical Association*, 94(447):847–854, Sep 1999.