DISTRIBUTED MAXIMUM LIKELIHOOD ESTIMATION FOR SENSOR NETWORKS

Doron Blatt and Alfred Hero

Department of EECS, University of Michigan, Ann Arbor, Michigan, USA
email address: dblatt@eecs.umich.edu, hero@eecs.umich.edu

ABSTRACT

The problem of finding the maximum likelihood estimator of a commonly observed model, based on data collected by a sensor network under power and bandwidth constraints is considered. In particular, a case where the sensors cannot fully share their data is treated. An iterative algorithm that relaxes the requirement of sharing all the data is given. The algorithm is based on a local Fisher scoring method and an iterative information sharing procedure. The case where the sensors share sub-optimal estimates is also analyzed. The asymptotic distribution of the estimates is derived and used to provide means of discrimination between estimates that are associated with different local maxima of the log-likelihood function. The results are validated by a simulation.

1. INTRODUCTION

The advent of a large number of applications for sensor networks has increased interest in the fields of distributed detection, estimation and quantization (see e.g. [1] and references therein). There have been two major streams in the research on distributed information gathering. The first, often called data fusion, uses heuristics in order to provide ad hoc methods for distributed information systems (see [2] and references therein). The other area of research uses information theory in order to gain insight on distributed systems, to derive bounds on their performance, and to construct algorithms for detection and estimation (see e.g. [3], [4], and [5]). Recently, Nowak [6] treated the problem of Maximum Likelihood (ML) estimation of the Gaussian mixture model by a sensor network and offered the Decentralized Expectation Maximization (EM) algorithm.

In the present paper, we adopt a different approach. We use asymptotic statistical theory in order to characterize certain aspects of the distributed system and to offer methods for performing estimation under power and bandwidth constraints.

The general setting considered in this paper is the following. A network of sensors is distributed in order to collect measurements of a common physical phenomenon. The data are collected for a parameter estimation task. This problem becomes trivial under any of the following conditions: (1) all the data can be shared by the sensors, or (2) a sufficient statistic is available and can be shared by the sensors. In these cases, maximum likelihood estimation can be performed and it is asymptotically optimal. However, if bandwidth and power constraints prevent sharing all the data and a sufficient statistic is not available, then questions arise as to how close we can get to optimal performance and by what means. For simplicity, we do not treat quantization issues. However, our results are extendable to quantized data and communication.

First we describe a simple information sharing method, which ensures that all sensors can compute the ML estimator (MLE) of the full data set collected by the sensor network. This can lead to a major reduction in the amount of transmitted information in the network without any loss of performance. Second, we present a sub-optimal reduced communication method in which each sensor computes a stationary point of its local likelihood function but not necessarily the MLE. These sub-optimal estimates are shared by the sensors. To aggregate these estimates, we apply an asymptotic theorem to approximate their distribution, which leads to a well posed Gaussian mixture framework. Finally, we validate and provide additional insight on the theoretical results by a simulation.

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

2. PROBLEM FORMULATION

Consider the following distributed estimation problem. A network of L sensors is geographically distributed in order to collect data for an estimation task. Each of these sensors collects independent $P \times 1$ random vectors $\mathbf{y}_{l,t}$, $t = 1, \ldots, n_l$; $l = 1, \ldots, L$, drawn independently from the same distribution with density $f(\mathbf{y}; \boldsymbol{\theta}^0)$. The $K \times 1$ vector parameter $\boldsymbol{\theta}^0$ is unknown but the density is known to lie in a parametric class $\{f(\mathbf{y}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. This scenario corresponds to L sensors that observe the same physical phenomenon through independent noisy channels.

Denote by $L(\mathbf{Y}_l; \boldsymbol{\theta}) = \frac{1}{n_l} \sum_{t=1}^{n_l} \log f(\mathbf{y}_{l,t}; \boldsymbol{\theta})$, where $\mathbf{Y}_l = [\mathbf{y}_{l,1}\, \mathbf{y}_{l,2}\, \ldots\, \mathbf{y}_{l,n_l}]$, the **local likelihood** of sensor l, which is the normalized log-likelihood function of the measurements available to sensor l. It is known that under some regularity conditions on $f(\mathbf{y}; \boldsymbol{\theta})$: as $n_l \to \infty$, $\frac{1}{n_l} \sum_{t=1}^{n_l} \log f(\mathbf{y}_{l,t}; \boldsymbol{\theta}) \to \mathrm{E}\{\log f(\mathbf{y}; \boldsymbol{\theta})\}$ uniformly in $\boldsymbol{\Theta}$ for almost every sequence $\{\mathbf{y}_{l,t}\}$. Here and in the sequel, $\mathrm{E}\{\cdot\}$ denotes the expectation with respect to the true density $f(\mathbf{y}, \boldsymbol{\theta}^0)$. Therefore,

$$\mathrm{E}\{\log f(\mathbf{y}; \boldsymbol{\theta})\} = \int \log\left(f(\mathbf{y}; \boldsymbol{\theta})\right) f(\mathbf{y}, \boldsymbol{\theta}^0) \mathrm{d}\mathbf{y} \stackrel{\triangle}{=} a(\boldsymbol{\theta}^0, \boldsymbol{\theta}) \ ,$$

which will be called the ambiguity function.

Denote by $\mathbf{Y} = [\mathbf{Y}_1\, \mathbf{Y}_2\, \ldots\, \mathbf{Y}_L]$ the **full data set** that is collected by the sensor network, and by $L(\mathbf{Y}; \boldsymbol{\theta}) = \frac{1}{L} \sum_{l=1}^{L} L(\mathbf{Y}_l; \boldsymbol{\theta})$ the normalized log-likelihood function associated with these data, which will be called the **global likelihood**.

Denote by $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{Y}; \boldsymbol{\theta})$ the global maximum likelihood estimator (GMLE), which is the MLE based on the global likelihood, and by $\widehat{\boldsymbol{\theta}}_l = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{Y}_l; \boldsymbol{\theta})$ the l^{th} local MLE (LMLE) which is based on the data of sensor l.

In many practical scenarios, bandwidth and power constraints prevent the sensors from sharing all of their data. Instead, only

partial information can be shared.

3. FISHER SCORING WITH ITERATIVE INFORMATION SHARING

When the maximization problem required for finding the MLE is intractable, iterative methods are often used. These methods generate a sequence $\{\boldsymbol{\theta}_i\}_{i\geq 1}$ which converges to a relative maximum of the log-likelihood function. If the log-likelihood function is not strictly convex over $\boldsymbol{\Theta}$, several initializations may be required in order to find the MLE.

In the context of sensor networks these methods can be used to iteratively find the GMLE without sharing the full data set. One possible method is Fisher scoring [7], in which $\boldsymbol{\theta}_i$ is updated by

$$\begin{aligned}\boldsymbol{\theta}_{i+1} &= \boldsymbol{\theta}_i + \mathbf{I}^{-1}(\boldsymbol{\theta}_i)\nabla L(\mathbf{Y};\boldsymbol{\theta}_i) \\ &= \boldsymbol{\theta}_i + \mathbf{I}^{-1}(\boldsymbol{\theta}_i)\frac{1}{L}\sum_{l=1}^{L}\nabla L(\mathbf{Y}_l;\boldsymbol{\theta}_i)\ . \end{aligned} \quad (1)$$

where $\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\left\{\nabla^2 L(\mathbf{Y};\boldsymbol{\theta})\right\}$, and for any function $g(\boldsymbol{\theta})$, $\nabla g(\boldsymbol{\theta})$ and $\nabla^2 g(\boldsymbol{\theta})$ denotes the vector of partial derivatives and the Hessian matrix of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, respectively. We call $\nabla L(\mathbf{Y};\boldsymbol{\theta}_i)$ the **global score function**. The second equality in (1) follows from the independent sensors assumption. Under broad conditions $\{\boldsymbol{\theta}_i\}_{i\geq 1}$ converges to a relative maximum of the global likelihood.

Without information sharing, each sensor can only implement the local updates

$$\boldsymbol{\theta}_{l,i+1} = \boldsymbol{\theta}_{l,i} + \mathbf{I}^{-1}(\boldsymbol{\theta}_{l,i})\nabla L(\mathbf{Y}_l;\boldsymbol{\theta}_{l,i}), \quad l = 1,\ldots,L\ , \quad (2)$$

which will converge to a relative maximum of the local likelihood. $\nabla L(\mathbf{Y}_l;\boldsymbol{\theta}_i)$ is called the **local score function** of sensor l. If sufficient bandwidth is available, a simple information sharing protocol can be deployed to perform the iterations in (1) in a distributed manner and to ensure that the sequence $\{\boldsymbol{\theta}_i\}_{i\geq 1}$ converges to a relative maximum of the global likelihood. Similar to the distributed implementation of the EM algorithm in [6], messages are cyclicly passed between sensors in sequential order. In the first cycle, $\boldsymbol{\theta}_1$ is shared by the sensors. Given $\boldsymbol{\theta}_i$, $\sum_{l=1}^{L}\nabla L(\mathbf{Y}_l;\boldsymbol{\theta}_i)$ is summed cumulatively as each sensor receives the running sum from the previous sensor, adds its local contribution, and sends the updated sum to the next sensor. In an additional cycle, the last sensor shares the total sum with the network and each sensor computes $\boldsymbol{\theta}_{i+1}$ according to (1). The procedure ends either when the relative change in the estimator's value is small or the sum of the score function values is close enough to zero. As mentioned before, several initializations may be required in order to find the GMLE.

4. AGGREGATION OF SUBOPTIMAL LOCAL ESTIMATES

As the above information sharing protocol requires sharing data at each iteration of (1), it may not be practical. Here we consider a scenario where the information is shared only once, after the convergence of each local search in (2). Denote by $\widetilde{\boldsymbol{\theta}}_l$ the relative maximum of $L(\mathbf{Y}_l;\boldsymbol{\theta})$ that the iterations of the l^{th} sensor converged to. Note that $\widetilde{\boldsymbol{\theta}}_l$ does not necessarily equal the LMLE $\widehat{\boldsymbol{\theta}}_l$. Denote the information shared by the sensors by $\{\widetilde{\boldsymbol{\theta}}_l, \boldsymbol{\eta}_l\}_{l=1}^{L}$, where $\boldsymbol{\eta}_l$ is an additional statistic shared by the sensors (e.g. $L(\mathbf{Y}_l;\widetilde{\boldsymbol{\theta}}_l)$ or the empty set). The question then arises as to how to treat the large number of estimates, some of which may correspond to convergence to the highest relative maximum and some to other relative maxima. Two questions will be answered below: (1) given the collection of estimates $\{\widetilde{\boldsymbol{\theta}}_l\}_{l=1}^{L}$, how to aggregate them and find an estimate for $\boldsymbol{\theta}^0$? and (2), how to use additional statistics of the data to improve and simplify the estimation of $\boldsymbol{\theta}^0$?

The approximation of the asymptotic distribution of a suboptimal estimator will lead to an estimate for $\boldsymbol{\theta}^0$. To this end we make the following assumption. Assume that $a(\boldsymbol{\theta}^0,\boldsymbol{\theta})$ has a finite number of relative maxima and minima with negative definite and positive definite Hessian matrices, respectively. Assume that all sensors collect the same number of samples and denote it by n. Then, the mathematical treatment is the same for all sensors and hence the subscript l is omitted. Denote the relative maxima of $a(\boldsymbol{\theta}^0,\boldsymbol{\theta})$ by $\boldsymbol{\theta}^m$, $m = 0,\ldots,M$.

Theorem 1 *Under the above assumption, $\exists N$ such that $\forall n > N$, $L(\mathbf{Y};\boldsymbol{\theta})$ has $M+1$ local maxima $w.p.1$, and the location of these relative maxima are strongly consistent estimates for $\boldsymbol{\theta}^m$, $m = 0,\ldots,M$.*

Proofs for all theorems are given in [8]. In order to derive the asymptotic distribution of an estimator associated with a relative maximum, consider the following setting. Let $\boldsymbol{\Theta}^m$ be a closed neighborhood of $\boldsymbol{\theta}^m$, in which $\boldsymbol{\theta}^m$ is the highest relative maximum of $a(\boldsymbol{\theta}^0,\boldsymbol{\theta})$. Define the m'th local MLE by $\widehat{\boldsymbol{\theta}}^m = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^m} L(\mathbf{Y};\boldsymbol{\theta})$. Define the matrices $\mathbf{A}(\boldsymbol{\theta}) = \mathrm{E}\left\{\nabla^2 \log f(\mathbf{y};\boldsymbol{\theta})\right\}$ and $\mathbf{B}(\boldsymbol{\theta}) = \mathrm{E}\left\{\nabla \log f(\mathbf{y};\boldsymbol{\theta})\cdot\nabla^T \log f(\mathbf{y};\boldsymbol{\theta})\right\}$, and when the inverse exists, the matrix

$$\mathbf{C}(\boldsymbol{\theta}) = \mathbf{A}^{-1}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\mathbf{A}^{-1}(\boldsymbol{\theta})\ . \quad (3)$$

Theorem 2 *Under the assumptions made above, for all m: (1) There exist a measurable $\widehat{\boldsymbol{\theta}}^m$ for all n, (2) $\widehat{\boldsymbol{\theta}}^m \overset{a.s.}{\to} \boldsymbol{\theta}^m$ as $n\to\infty$, and (3) $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^m - \boldsymbol{\theta}^m\right) \overset{D}{\to} N\left(\mathbf{0}_{K\times 1}, \mathbf{C}(\boldsymbol{\theta}^m)\right)$.*

For $\widehat{\boldsymbol{\theta}}^0 = \widehat{\boldsymbol{\theta}}$, Theorem 2 is the standard existence, consistency, and asymptotic Gaussian distribution of the MLE, with $\mathbf{C}(\boldsymbol{\theta}^0)$ equals to the inverse of the Fisher information matrix (FIM).

Furthermore, consider any $Q\times 1$ vector valued function $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$, which is bounded and twice differentiable with respect to $\boldsymbol{\theta}$ with bounded derivatives. Define the vectors $\mathbf{h}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{e}(\mathbf{y}_t,\boldsymbol{\theta})$ and $\mathbf{h}(\boldsymbol{\theta}) = \mathrm{E}\{\mathbf{e}(\mathbf{y},\boldsymbol{\theta})\}$ and the $Q\times K$ partial derivatives matrix $[\mathbf{H}(\boldsymbol{\theta})]_{q,k} = \mathrm{E}\{\partial e_q(\mathbf{y},\boldsymbol{\theta})/\partial\theta_k\}$. When the expectation exists, define the $(Q+K)\times(Q+K)$ matrix

$$\mathbf{W}(\boldsymbol{\theta}) = \mathrm{E}\left\{\begin{bmatrix} \boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla\log f(\mathbf{y};\boldsymbol{\theta}) \\ \mathbf{e}(\mathbf{y},\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla\log f(\mathbf{y};\boldsymbol{\theta})\end{bmatrix} \times \begin{bmatrix} \boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla\log f(\mathbf{y};\boldsymbol{\theta}) \\ \mathbf{e}(\mathbf{y},\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\boldsymbol{A}^{-1}(\boldsymbol{\theta})\nabla\log f(\mathbf{y};\boldsymbol{\theta})\end{bmatrix}^T\right\}.$$

Assume that $\mathbf{W}(\boldsymbol{\theta}^m)$ is nonsingular for all m. In practice, this assumption is satisfied by an appropriate choice of $\mathbf{e}(\mathbf{y},\boldsymbol{\theta})$.

Theorem 3 *Under the assumptions made above, for all m,*

$$\sqrt{n}\begin{bmatrix}\widehat{\boldsymbol{\theta}}^m - \boldsymbol{\theta}^m \\ \mathbf{h}_n(\widehat{\boldsymbol{\theta}}^m) - \mathbf{h}(\boldsymbol{\theta}^m)\end{bmatrix} \overset{D}{\to} N\left(\mathbf{0}_{(K+Q)\times 1}, \mathbf{W}(\boldsymbol{\theta}^m)\right)\ .$$

Theorems 1- 3 provide the means for approximating the asymptotic density of $\widetilde{\boldsymbol{\theta}}_l$, denoted by $f_{\widetilde{\theta}_l}(\boldsymbol{\theta};\boldsymbol{\theta}^0)$, and the asymptotic joint density of $\widetilde{\boldsymbol{\theta}}_l$ and a statistic that is based on the data and the estimator, denoted by $f_{\widetilde{\boldsymbol{\theta}}_l,\mathbf{h}_n(\widetilde{\theta}_l)}(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\theta}^0)$. If the iterative local search in (2) is certain to find a relative maximum of the local likelihood, then Theorem 1 guarantees that for sufficiently large n the final estimate will be in the vicinity of one of the $M+1$ relative maxima of $a(\boldsymbol{\theta}^0,\boldsymbol{\theta})$ $w.p.1$. Then, defining D^m as the event that the estimator $\widetilde{\boldsymbol{\theta}}_n$ is in $\boldsymbol{\Theta}^m$ and denoting its probability by $\mathbb{P}_n(D^m;\boldsymbol{\theta}^0)$, we obtain that for sufficiently large n, $D^{m_1}\bigcap D^{m_2}=\emptyset$ and $\mathbb{P}(\bigcup_{m=0}^{M}D^m)=1$. In general, $\mathbb{P}_n(D^m;\boldsymbol{\theta}^0)$ depends on n, the true parameter and the initialization method.

Corollary 1 *Under the assumptions made above, for sufficiently large n,*

$$f_{\widetilde{\boldsymbol{\theta}}_l}(\boldsymbol{\theta};\boldsymbol{\theta}^0)\approx\sum_{m=0}^{M}\frac{\mathbb{P}_n(D^m;\boldsymbol{\theta}^0)}{(2\pi)^{K/2}\sqrt{\det(1/n\mathbf{C}(\boldsymbol{\theta}^m))}}$$
$$\exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^m)^T\mathbf{C}^{-1}(\boldsymbol{\theta}^m)(\boldsymbol{\theta}-\boldsymbol{\theta}^m)\right\}\ ,$$

where $\mathbf{C}(\boldsymbol{\theta}^0)=-\mathbf{A}(\boldsymbol{\theta}^0)=\mathbf{B}(\boldsymbol{\theta}^0)$ *and*

$$f_{\widetilde{\boldsymbol{\theta}}_l,\mathbf{h}_n(\widetilde{\theta}_l)}(\boldsymbol{\theta},\mathbf{x};\boldsymbol{\theta}^0)\approx\sum_{m=0}^{M}\frac{\mathbb{P}_n(D^m;\boldsymbol{\theta}^0)}{(2\pi)^{K/2}\sqrt{\det(1/n\mathbf{W}(\boldsymbol{\theta}^m))}}\times$$
$$\exp\left\{-\frac{n}{2}\begin{bmatrix}\boldsymbol{\theta}-\boldsymbol{\theta}^m\\ \mathbf{x}-\mathbf{h}(\boldsymbol{\theta}^m)\end{bmatrix}^T\mathbf{W}^{-1}(\boldsymbol{\theta}^m)\begin{bmatrix}\boldsymbol{\theta}-\boldsymbol{\theta}^m\\ \mathbf{x}-\mathbf{h}(\boldsymbol{\theta}^m)\end{bmatrix}\right\}\ .$$

When the information sharing makes all the local estimates available, Corollary 1 provides the means to find a good approximation to the GMLE $\widehat{\boldsymbol{\theta}}$ through a well-posed Gaussian mixture problem. The theory asserts that these sub-optimal estimates are drawn from a distribution, which is approximately a multivariate Gaussian Mixture. Furthermore, the cluster corresponding to local estimates which are close to the highest maximum of the global likelihood has the property that its covariance matrix is close to the inverse of the FIM evaluated at the mean of this cluster of estimates. This property can be used to discriminate between relative maxima.

First, the number of components, the mean vectors and the covariance matrices, of the multivariate Gaussian mixture distribution of $\{\widetilde{\boldsymbol{\theta}}_l\}_{l=1}^{L}$ are estimated. The state-of-the-art estimator for mixture models is the CEM given in [9]. The estimated mean vectors serve as candidates for the final estimate and the estimated covariance matrices provide the means to find the component that corresponds to the GMLE. Explicitly, for each component the distance (e.g. Frobenius norm) between the estimated covariance and the inverse of the FIM calculated at the point of the mean of this component is computed and the mean of the component with the smallest distance is chosen as the final estimate. If the FIM cannot be computed analytically, it needs to be computed by numerical integration. As will be shown in section 5, this method provides reliable discrimination between estimates that are associated with the global maximum and estimates that are associated with local maxima without the need to cluster each estimate separately.

If additional information is shared by the sensors, it can be used to improve the clustering and to provide additional discrimination. For example, if the data shared by the sensors is the set $\{\widetilde{\boldsymbol{\theta}}_l, L(\mathbf{Y}_l;\widetilde{\boldsymbol{\theta}}_l)\}_{l=1}^{L}$, then the final estimate can be the mean of the cluster of estimates with the highest average log-likelihood value.

5. SIMULATION RESULTS

We simulated a network of L 2D position estimating sensors and evaluated two cases: (1) The GMLE under the iterative information sharing discussed in section 3, and (2) The partial information sharing discussed in section 4. In the simulation, $n=50$ samples for each of the L sensors were generated according to the following bivariate Gaussian mixture density $f(\mathbf{y};\boldsymbol{\theta})=\sum_{j=1}^{2}\alpha_j f(\mathbf{y};\boldsymbol{\mu}_j)$, where $f(\mathbf{y};\boldsymbol{\mu}_j)$ is a bivariate Gaussian density with unknown mean $\boldsymbol{\mu}_j=[\mu_{j1}\ \mu_{j2}]^T$, and known covariance matrix $\mathbf{R}=0.2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and where $\mathbf{y}=[y_1\ y_2]^T$. The known mixing probabilities are $\alpha_1=1-\alpha_2=0.4$. The parameter vector $\boldsymbol{\theta}=[\mu_{11}\,\mu_{12}\,\mu_{21}\,\mu_{22}]^T$ is known a-priori to lie in $\boldsymbol{\Theta}=[0,3]\times[0,3]\times[0,3]\times[0,3]$. We chose a scenario in which the number of samples is small in order to demonstrate that our method is not restricted to the asymptotic regime. Note that the Gaussian mixture model of the sensors' data has nothing to do with the Gaussian mixture model of Theorem 3, which is an asymptotic distribution for the aggregation of the sub-optimal estimates.

Each sensor uses local Fisher scoring via (2) to compute its estimate, where each starting point $\boldsymbol{\theta}_1$ is generated randomly, according to a uniform distribution on $\boldsymbol{\Theta}$. Using this method, we observed that about half of the sensors found an estimate which is in the vicinity of the true parameter. The other half stagnated at a local maximum.

The ambiguity function has two maxima in $\boldsymbol{\Theta}$, one at the true parameter $\boldsymbol{\theta}^0=[1\,2\,2\,1]$ and one at $\boldsymbol{\theta}^1=[2.05\,0.95\,1.08\,1.92]^T$. Therefore, Corollary 1 asserts that the aggregate distribution of the estimates $\{\widetilde{\boldsymbol{\theta}}_l\}_{l=1}^{L}$ is approximately a two component 4D multivariate Gaussian mixture, where the vector means of the two components are the locations of the two maxima of the ambiguity function and the covariance matrices are $\mathbf{C}(\boldsymbol{\theta}^0)$ and $\mathbf{C}(\boldsymbol{\theta}^1)$ given in (3). A realization of 4D estimates generated by $L=200$ sensors is presented in Fig. 1. Each sub-figure corresponds to a projection of the 4D estimates onto a 2D subspace, which is either the first two or the last two coordinates. This collection of estimates is used to find a final estimate via the aggregation method described earlier. Clearly, there is a better fit between the estimated covariance and the computed FIM at the cluster that corresponds to estimates in the vicinity of the highest relative maximum of the ambiguity function. The mean of this cluster of estimates is used as the approximation to the GMLE.

The performance of this method was evaluated as the number of sensors L increases. The results are summarized in Fig. 2. Before estimating the averaged MSE, the cases in which a clustering error has occurred were excluded. When $L>100$ a clustering error occurred in less then one percent of the trials. First, the performance of the GMLE and the Cramer Rao bound (CRB) are presented as a benchmark. The performance of the GMLE using iterative information sharing corresponds to the Fisher scoring method presented in section 3. This method attains the (CRB) for $L>10$. The average of LMLEs corresponds to the performance of a network of sensors that perform a global maximization and, in contrast to our aggregation method, always find the LMLE. However, this network performs a crude aggregation rule that simply averages the local estimates. The aggregation method, given by clustering estimates according to the Gaussian mixture of Corollary 1 outperforms this crude averaging, even though the quality of the individual local estimates is worse. In other words, we are able to compensate for the sub-optimality of the individual sensors

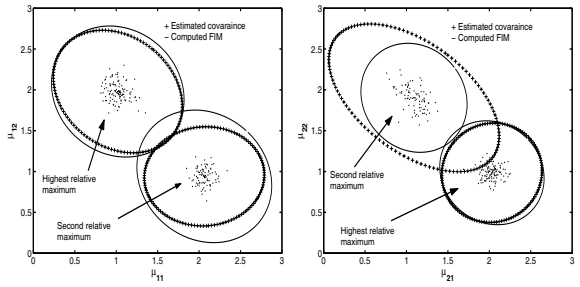

Fig. 1. Discrimination according to covariance.

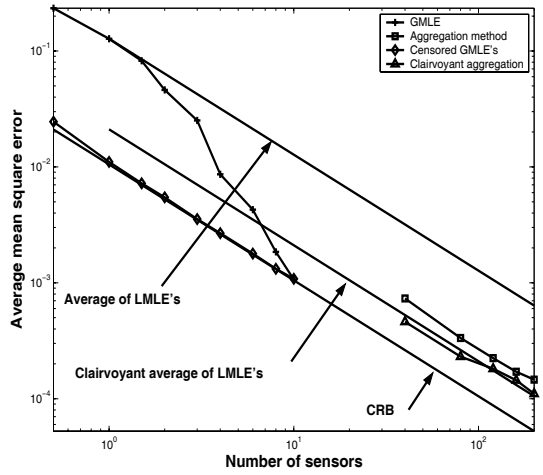

Fig. 2. Performance evaluation.

by additional processing at the aggregation stage. In addition, we added the performance of a clairvoyant aggregation method which averages only those estimates that are close to the global maximum. The degradation in performance of our method in comparison to this clairvoyant method is the price we pay for clustering.

Also indicated is the performance of the censored GMLEs, which is the performance of the GMLE in cases where it found a maximum at the vicinity of the true parameter. The clairvoyant average of LMLE's corresponds to a system with half the number of sensors but with sensors that always find the closest maximum to the true parameter. We conclude that our method not only compensates for the sub-optimality of the sensors, but it performs almost as if the individual sensors were operating in the asymptotic regime, in which case the LMLE's themselves are indeed sufficient statistics.

If the values of the local likelihood at $\widetilde{\boldsymbol{\theta}}_l$ are shared among the sensors, the second part of Corollary 1 can be invoked. The values of the log-likelihood function at the global and local maxima are nearly identical and it is not clear from the histogram in Fig. 3 that there are two separable components. However, the coupling of the log-likelihood values with the local estimates improves the discrimination between the relative maxima. As described in section 4, the discrimination between the components of the global and local maxima can be done by using the estimated means alone, without the need to calculate the FIM. Our simulations showed that this simple strategy provides similar results as the previous one.

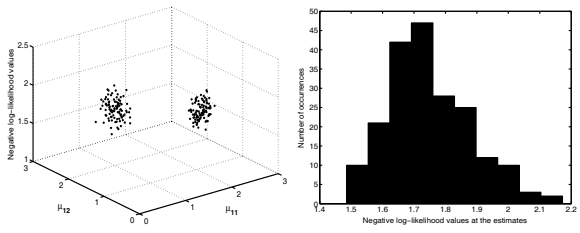

Fig. 3. Incorporating the log-likelihood values.

The cost of this simplification is increased bandwidth, namely, the transmission of additional information, the scalars $L(\mathbf{Y}_l; \widetilde{\boldsymbol{\theta}}_l)$.

6. CONCLUDING REMARKS

The problem of finding the MLE based on data collected by a sensor network under power and bandwidth constraints was considered. An iterative information sharing protocol which is based on the Fisher scoring method was given as a method for finding the global MLE without sharing the full data set. For cases in which iterative information sharing is prohibited by a bandwidth constraint, an alternative method was given. Instead of iterative information sharing, each sensor finds a sub-optimal estimate based on its local data and shares this estimate once with the other sensors. An asymptotic theorem was applied to approximate the distribution of these sub-optimal estimates which provided the means for aggregating these estimates into a final global estimate. The aggregation method compensates for the sub-optimality of the sensors.

7. REFERENCES

[1] S. Kumar, F. Zhao, and D. Shepherd editors, "Special issue on collaborative information processing," *IEEE Signal Processing Magazine*, vol. 19, no. 2, March 2002.

[2] P. K. Varshney, *Distributed Detection and Data Fusion*, Springer-Verlag, 1997.

[3] P. Ishwar, R. Puri, S. S. Pradhan, and K. Ramchandran, "On compression for robust estimation in sensor networks," *Proc.of International Symposium on Information Theory (ISIT), Yokohama, Japan*, June 2003.

[4] T.S. Han and S. Amari, "Parameter estimation with multiterminal data compression," *IEEE transactions on Information Theory*, vol. 41, no. 6, pp. 1802 –1833, Nov. 1995.

[5] J. F. Chamberland and V.V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407 –416, Feb. 2003.

[6] R. D. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE transactions on signal processing*, vol. 51, no. 8, pp. 2245–2253, August 2003.

[7] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, 1997.

[8] D. Blatt and A. Hero, "Distributed maximum likelihood for sensor networks," *In preparation*.

[9] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixute models," *IEEE Trans on Pattern Anal and Machine Intelligence*, vol. 24, pp. 381–396, March 2002.