ON SEQUENTIAL ESTIMATION OF LINEAR MODELS FROM DATA WITH CORRELATED NOISE

Yunlong Wang and Petar M. Djurić

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794
Email: {yunlong.wang, petar.djuric}@stonybrook.edu

ABSTRACT

In this paper, we consider the problem of Bayesian sequential estimation on a set of time invariant parameters. At every time instant, a new observation through a linear model is obtained where the observations are distorted by spatially correlated noise with unknown covariance, whereas in time, the noise samples are independent and identically distributed. We derive the joint posterior of the parameters of interest and the covariance, and we propose several approximations to make the Bayesian estimation tractable. Then we propose a method for forming a pseudo posterior, which is suitable for settings where estimation over networks is applied. By computer simulations, we demonstrate that the Kullback–Leibler divergence between the pseudo posterior and a posterior obtained from a known covariance decreases as the acquisition of new observations continues. We also provide computer simulations that compare the proposed method with the least squares method.

***Index Terms*—** Bayesian inference, distributed estimation, unknown covariance, pseudo posterior

1. INTRODUCTION

Estimation over cooperative networks has been widely studied in the literature (e.g., [1, 2]), where the agents estimate the state of nature in a distributed manner by exchanging information with neighbors. In [3, 4, 5], we address consensus-based distributed estimation of linear models within the Bayesian framework.

In addressing consensus-based estimation methods over networks, it is important to reformulate the formation of the optimal posterior to be a function of summation of certain statistics. In [4] and [6], it is shown that by using this strategy, the original problem can be converted to a problem of distributed summation, which is suitable for the consensus method [7]. In [8], the authors consider a problem where the observation noises are spatially correlated with known covariance matrix. In this paper, we are laying the grounds for distributed estimation to scenarios where the observation noise is spatially correlated with unknown covariance, whereas in time, the noise is independent and identically distributed. We focus on the required processing by an agent at every time step and show why it cannot obtain the exact posteriors of the unknowns. Then we resort to a suboptimal approach that can be used to approximate the individual agent's belief to a suitable form. This belief can then readily be used in settings of distributed estimation.

More specifically, we study sequential estimation where at every time instant an agent gets observations from several sensors about a vector of time invariant parameters through a linear model. In order to get the posterior of the parameter of interests, the agent needs to marginalize out the unknown covariance. The marginalization, however, is computationally intractable. Therefore, we define a pseudo posterior as an approximation of the posterior and propose a method for the agent to approximately reach the optimal Bayesian result. This approach lends itself readily to processing over networks of agents.

The paper is organized as follows. In the next section we state the problem. In Section 3, we present the mathematical development of the proposed method. With the result in Section 3, we propose the method for estimation in Section 4. Section 5 provides simulation results, and Section 6 contains conclusions.

2. PROBLEM FORMULATION

We address sequential Bayesian estimation of a vector of linear parameters $\boldsymbol{\theta} \in \mathbb{R}^{K\times 1}$ by a single agent. The solution of this problem is a key to resolving the problem of distributed Bayesian estimation, where each agent receives measurements from its sensors, with the measurements being distorted by time independent but spatially correlated noise. In Fig. 1, we display a general scenario of distributed estimation, where each agent has M sensors providing the

This work was supported by NSF under Awards CCF-1320626 and ECCS-1346854.

agent with spatially correlated observations. The agents exchange information for cooperative estimation of the unknowns of interest.

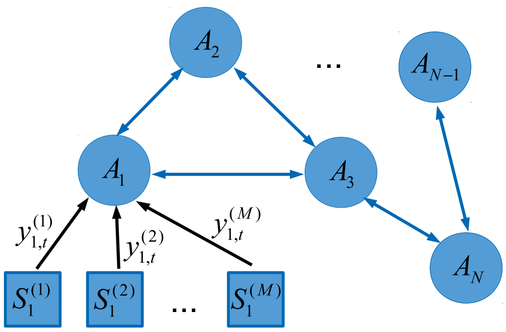

Fig. 1. A schematic diagram of a distributed system, where A_n denotes the nth agent, $S_n^{(m)}$ represents the mth sensor of agent A_n, and $y_{n,t}^{(m)}$ denotes the measurements provided from $S_n^{(m)}$ to A_n at time instant t.

Consider that at each time instant $t \in \mathbb{N}^+$, an agent receives data $\mathbf{y}_t$ generated by a linear model of the form

$$\mathbf{y}_t = \mathbf{H}_t\boldsymbol{\theta} + \mathbf{w}_t, \tag{1}$$

where $\mathbf{H}_t \in \mathbb{R}^{M\times K}$ is a matrix known by the agent, and $\mathbf{w}_t$ denotes the observation noise modeled as a Gaussian random vector with zero-mean and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$, which is time invariant but unknown. It is also assumed that $M \geq K$ and the observation noise $\mathbf{w}_t$ is white in time.

We use the model in (1) and define the agent's prior on $\boldsymbol{\Sigma}$ as an inverse Wishart distribution, $\mathcal{W}^{-1}(\boldsymbol{\Lambda}_0, \nu)$, given by

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\nu+M+1}{2}} \exp\left(-\frac{1}{2}tr(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1})\right), \tag{2}$$

where $\boldsymbol{\Lambda}_0$ is a scale matrix, ν represents degrees of freedom, and $tr(\cdot)$ denotes the trace of the matrix inside the parentheses. Let the prior of $\boldsymbol{\theta}$ be Gaussian denoted by

$$p(\boldsymbol{\theta}|\boldsymbol{\Sigma}) \propto |\mathbf{C}_0^{-1}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{C}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right), \tag{3}$$

where $^\top$ denotes matrix transpose. We set $\mathbf{C}_0^{-1} = \mathbf{H}_0^\top\boldsymbol{\Sigma}^{-1}\mathbf{H}_0$.

From the Bayes' rule, the posteriors of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ can be obtained by

$$p\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathcal{I}_t\right) \quad \propto \quad p\left(\mathcal{I}_t|\boldsymbol{\theta}, \boldsymbol{\Sigma}\right)p\left(\boldsymbol{\theta}|\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\Sigma}\right), \tag{4}$$

where $\mathcal{I}_t$ refers to all the information up to time t, which includes $\mathbf{y}_\tau$, $\mathbf{H}_\tau$, for all $\tau \in \{1, 2, \cdots, t\}$. We propose that the agent formulates its belief about $\boldsymbol{\theta}$ as the marginalized posterior of $\boldsymbol{\theta}$ given by,

$$\begin{aligned}\beta_t &= p(\boldsymbol{\theta}|\mathcal{I}_t)\\ &= \int_{\boldsymbol{\Sigma}\succ\mathbf{0}} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathcal{I}_t)d\boldsymbol{\Sigma},\end{aligned} \tag{5}$$

where $\boldsymbol{\Sigma} \succ \mathbf{0}$ denotes that $\boldsymbol{\Sigma}$ is positive definite.

We also define a benchmark for this problem. We assume that there exists a genie agent which knows both $\mathcal{I}_t$ and $\boldsymbol{\Sigma}$. Its belief is given by,

$$\beta_t^{(opt)} \quad = \quad p(\boldsymbol{\theta}|\mathcal{I}_t, \boldsymbol{\Sigma}). \tag{6}$$

In this work, our aim is to derive expressions for sequential update of the posterior β_t and the estimate of $\boldsymbol{\theta}$.

3. MATHEMATICAL DEVELOPMENT OF THE PROPOSED METHOD

In this section, we first derive the posterior held by the agent and then we propose the approximations to make the problem tractable for distributed estimation of $\boldsymbol{\theta}$.

3.1. Posterior of individual agent

Using the model in Section 2, the expression (1) suggests that the likelihood of the data at time t can be written by

$$p(\mathcal{I}_t|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{t}{2}} \exp\left(-\frac{1}{2}\sum_{\tau=1}^{t}\mathbf{w}_\tau^\top\boldsymbol{\Sigma}^{-1}\mathbf{w}_\tau\right) \tag{7}$$

with $\mathbf{w}_\tau = \mathbf{y}_\tau - \mathbf{H}_\tau\boldsymbol{\theta}$ denoting the observation noise.

Using the Bayes' rule, one can show that the posterior of $\boldsymbol{\theta}$ is given by

$$\begin{aligned}p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathcal{I}_t) &\propto p(\mathcal{I}_t|\boldsymbol{\theta}, \boldsymbol{\Sigma})p(\boldsymbol{\theta}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma})\\ &= |\boldsymbol{\Sigma}|^{-\frac{\nu+t+M+1}{2}}|\mathbf{C}_0|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}L_t\right),\end{aligned} \tag{8}$$

where L_t is defined by

$$\begin{aligned}L_t &= \sum_{\tau=1}^{t}\mathbf{w}_\tau^\top\boldsymbol{\Sigma}^{-1}\mathbf{w}_\tau + tr(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1})\\ &\quad +(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{C}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0).\end{aligned} \tag{9}$$

We show in the appendix that L_t can be reformulated into the following quadratic form:

$$L_t \quad = \quad (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}_t)^\top\mathbf{C}_t^{-1}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}_t) + S_t \tag{10}$$

with $\mathbf{C}_t$ and $\widetilde{\boldsymbol{\theta}}_t$ given by

$$\mathbf{C}_t = \left(\sum_{\tau=0}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{H}_\tau\right)^{-1}, \quad (11)$$

$$\widetilde{\boldsymbol{\theta}}_t = \mathbf{C}_t^{-1} \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau + \mathbf{C}_0^{-1}\boldsymbol{\theta}_0\right), \quad (12)$$

where S_t is defined by

$$S_t = (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t)^\top (\mathbf{C}_0 + \mathbf{M}_t)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t) + tr(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) + \sum_{\tau=1}^{t} (\mathbf{y}_\tau - \mathbf{H}_\tau \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\tau - \mathbf{H}_\tau \boldsymbol{\mu}_t). \quad (13)$$

In (13), $\boldsymbol{\mu}_t$ represents the maximum likelihood estimate of $\boldsymbol{\theta}$ in the form of

$$\boldsymbol{\mu}_t = \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{H}_\tau\right)^{-1} \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau\right), \quad (14)$$

and $\mathbf{M}_t$ denotes the covariance matrix of $\boldsymbol{\mu}_t$ defined by,

$$\mathbf{M}_t = \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{H}_\tau\right)^{-1}. \quad (15)$$

We write $p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathcal{I}_t) = p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathcal{I}_t)p(\boldsymbol{\Sigma}|\mathcal{I}_t)$, and from the expression of L_t in (10), we can get the optimal belief held by the genie agent $\beta_t^{(opt)} = p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathcal{I}_t)$. The belief is expressed by a multivariate Gaussian distribution given by

$$\beta_t^{(opt)} = \mathcal{N}\left(\widetilde{\boldsymbol{\theta}}_t, \mathbf{C}_t\right). \quad (16)$$

By using (8) and (10), we can show that after integrating out $\boldsymbol{\theta}$, the marginalized posterior of $\boldsymbol{\Sigma}$ has the following form:

$$p(\boldsymbol{\Sigma}|\mathcal{I}_t) \propto |\mathbf{C}_0 \mathbf{C}_t^{-1}|^{-\frac{1}{2}} |\boldsymbol{\Sigma}|^{-\frac{\nu+t+M+1}{2}} \exp\left(-\frac{1}{2} S_t\right), \quad (17)$$

where S_t is defined in (13).

3.2. Approximation of the posterior

The expression for $\boldsymbol{\mu}_t$ in (14) suggests that $\boldsymbol{\mu}_t$ is a function of $\boldsymbol{\Sigma}$, and thus, the posterior of $\boldsymbol{\Sigma}$ in (17) is not an inverse Wishart distribution. This means that the integration

$$p(\boldsymbol{\theta}|\mathcal{I}_t) = \int_{\boldsymbol{\Sigma} \succ \mathbf{0}} p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathcal{I}_t) p(\boldsymbol{\Sigma}|\mathcal{I}_t) d\boldsymbol{\Sigma}, \quad (18)$$

is not tractable.

However, from (17) and (13), $\boldsymbol{\Sigma}$ has an inverse Wishart distribution if the following three approximations are valid. First,

$$(\mathbf{C}_0 + \mathbf{M}_t)^{-1} \approx \mathbf{C}_0^{-1}, \quad (19)$$

which is valid for large t. The reason is that when t is large, the elements of $\mathbf{M}_t$ become very small in comparison to those of $\mathbf{C}_0$. Second,

$$|\mathbf{C}_0 \mathbf{C}_t^{-1}| \approx r_t, \ \forall t \in \mathbb{N}, \quad (20)$$

where r_t is approximately a constant. In the special case when the $\mathbf{H}_t$s are identical or proportional to each other for different ts, r_t is truly a constant.

The third approximation is that when t is large, the estimate of $\boldsymbol{\Sigma}$ held by an agent becomes close to the true value of $\boldsymbol{\Sigma}$. Then, in (14), we substitute $\boldsymbol{\Sigma}$ with $\widehat{\boldsymbol{\Sigma}}_{t-1}$, and we have the approximation $\boldsymbol{\mu}_t \approx \widehat{\boldsymbol{\mu}}_t$ where

$$\widehat{\boldsymbol{\mu}}_t = \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \widehat{\boldsymbol{\Sigma}}_{t-1}^{-1} \mathbf{H}_\tau\right)^{-1} \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \widehat{\boldsymbol{\Sigma}}_{t-1}^{-1} \mathbf{y}_\tau\right), \quad (21)$$

which means the agent can be viewed as if it knows the true value of $\boldsymbol{\Sigma}$ in calculating $\boldsymbol{\mu}_t$.

Even with these approximations, the joint posterior of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ is not a Normal Inverse Wishart distribution. Then we propose that the agent uses an additional approximation for generating its pseudo-posterior, i.e.,$p(\boldsymbol{\theta}|\mathcal{I}_t) \approx \widehat{\beta}_t$, where

$$\widehat{\beta}_t = p(\boldsymbol{\theta}|\mathcal{I}_t, \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}_t), \quad (22)$$

and where $\widehat{\boldsymbol{\Sigma}}_t$ denotes the latest estimate of $\boldsymbol{\Sigma}$ held by the agent at time instant t, which can be, e.g., be the maximum likelihood estimate. However, in this paper, we propose that the agent uses the MMSE estimate. This estimate is the mean of the marginalized posterior of $\boldsymbol{\Sigma}$, i.e.,

$$\widehat{\boldsymbol{\Sigma}}_{MMSE,t} = \int_{\boldsymbol{\Sigma} \succ \mathbf{0}} \boldsymbol{\Sigma}\, p(\boldsymbol{\Sigma}|\mathcal{I}_t)\, d\boldsymbol{\Sigma}. \quad (23)$$

With the above three assumptions, we can analytically solve the integration in (23) and obtain $\widehat{\boldsymbol{\Sigma}}_t$ by

$$\widehat{\boldsymbol{\Sigma}}_t = \frac{1}{\nu_t - M - 1} \Big(\sum_{\tau=1}^{t} (\mathbf{y}_\tau - \mathbf{H}_\tau \widehat{\boldsymbol{\mu}}_t)(\mathbf{y}_\tau - \mathbf{H}_\tau \widehat{\boldsymbol{\mu}}_t)^\top + \boldsymbol{\Lambda}_0 + \mathbf{H}_0 (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\mu}}_t)(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\mu}}_t)^\top \mathbf{H}_0^\top \Big), \quad (24)$$

where $\nu_t = \nu + t$ and $\widehat{\boldsymbol{\mu}}_t$ is defined in (21). Furthermore, this estimate approaches $\widehat{\boldsymbol{\Sigma}}_{MMSE,t}$ with time.

The expressions (16) and (22) show that the pseudo posterior of agent A_n of $\boldsymbol{\theta}$ has the following form:

$$\widehat{\beta}_t = \mathcal{N}\left(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{C}}_t\right), \quad (25)$$

where

$$\widehat{\mathbf{C}}_t = \left(\sum_{\tau=0}^{t} \mathbf{H}_\tau^\top \widehat{\boldsymbol{\Sigma}}_t^{-1} \mathbf{H}_\tau\right)^{-1}, \quad (26)$$

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\mathbf{C}}_t^{-1} \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \widehat{\boldsymbol{\Sigma}}_t^{-1} \mathbf{y}_\tau + \widehat{\mathbf{C}}_0^{-1}\boldsymbol{\theta}_0\right), \quad (27)$$

with $\widehat{\boldsymbol{\Sigma}}_t$ being defined in (24).

4. THE PROPOSED METHOD

In this section, we summarize the method for updating the pseudo posterior of $\boldsymbol{\theta}$ employed by the agent. At each time instant t, it implements the following steps:

Initialization: At $t = 0$, the agent forms its prior by (2) and (3), and initializes $\widehat{\boldsymbol{\Sigma}}_0$ by $\widehat{\boldsymbol{\Sigma}}_0 = \frac{\boldsymbol{\Lambda}_0}{\nu - M - 1}$. The steps below describe the tth recursion.

Step 1 The agent receives the data $\mathbf{y}_t$ and the regressors $\mathbf{H}_t$ from the sensors and calculates $\widehat{\boldsymbol{\mu}}_t$ by (21).

Step 2 With $\widehat{\boldsymbol{\mu}}_t$ and (24), the agent updates its estimate $\widehat{\boldsymbol{\Sigma}}_t$.

Step 3 The agent forms its pseudo posterior $\widehat{\beta}_t$ as a Gaussian distribution with a mean $\widehat{\boldsymbol{\theta}}_t$ and covariance $\widehat{\boldsymbol{\Sigma}}_t$ defined in (26) and (27).

We point out that since an agent approximates its belief by a Gaussian distribution, the distributed estimation problem can be converted to a distributed summation problem [6].

5. SIMULATION

In this section, we provide computer simulations that show the performance of our method in terms of convergence and numerical comparisons with the least squares (LS) method. We performed two experiments. In the first experiment, we implemented the proposed and the LS methods with identical random data in 1000 realizations. In each of the trials, $\boldsymbol{\theta} = [3, 3, 2, 2]^\top$, $M = 10$, $K = 4$, $t \in \{1, 2, \cdots, 200\}$. We set the elements of $\mathbf{H}_t \in \mathbb{R}^{M \times K}$ to be independent random variables uniformly distributed on $[3, 5]$. Also, in every trial, we drew $\boldsymbol{\Sigma}$ from its prior, an inverse Wishart distribution with $\boldsymbol{\Lambda}_0 = 10\mathbf{I}_M$ (with $\mathbf{I}_M \in \mathbb{R}^{M \times M}$ denoting the identity matrix) and $\nu = 12$. We also set $\boldsymbol{\theta}_0 = [0, 0, 0, 0]^\top$ and generated $\mathbf{H}_0$ in the same way as we did $\mathbf{H}_t$, $t > 0$.

As a performance metric for the different methods, we used the mean square deviation at time t, MSD(t), defined as the average value of $\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|^2$ over 1000 implementations. To demonstrate the advantage of the proposed method, we compared the MSD of the proposed method with that of the LS method and of the genie agent. The LS estimate of $\boldsymbol{\theta}$ at time instant t is given by

$$\widehat{\boldsymbol{\theta}}_t^{(LS)} = \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \mathbf{H}_\tau \right)^{-1} \left(\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \mathbf{y}_\tau \right). \quad (28)$$

The MSD of the genie agent served as a benchmark. We note that the estimate of $\boldsymbol{\theta}$ of the genie agent is $\widetilde{\boldsymbol{\theta}}_t$ given by (12).

In the second experiment, we repeated everything except that we generated $\boldsymbol{\Sigma}$ with another scale matrix $\mathbf{Q}^2$, where $\mathbf{Q} \in \mathbb{R}^{M \times M}$ was with elements that were independent random variables uniformly distributed on $[0, 5]$. The intention of this experiment was to show the performance of the proposed method with highly correlated data.

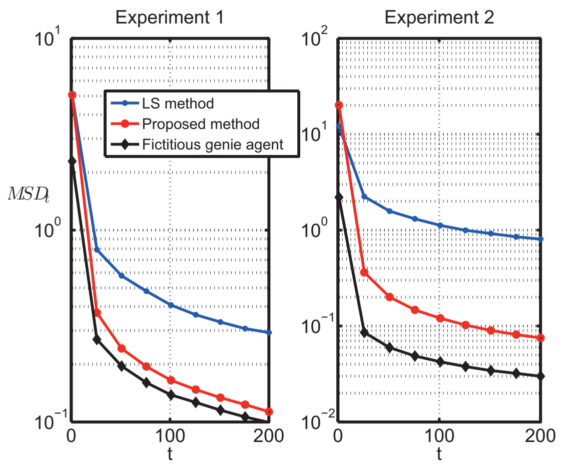

Fig. 2. Asymptotical performance of the proposed method.

The results of the two experiments are shown in Fig. 2 (on the left of experiment 1, and on the right of experiment 2), where we plotted the MSDs of the different methods. It can be seen that the proposed method shows a faster convergence than the LS method in terms of MSD. This difference is even more obvious when processing data with a higher correlation (as in experiment two).

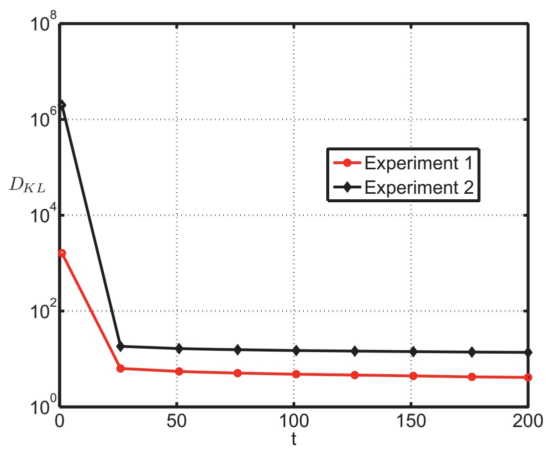

Fig. 3. The evolution of KL divergence.

In Fig. 3, we plotted the Kullback–Leibler (KL) divergence, $D_{KL}(\beta_t^{(opt)} || \widehat{\beta}_t))$, between the belief of the genie agent and the belief of the agent that employs the proposed method.

From the figure, it can be seen that the KL divergence keeps decreasing as time evolves, but it does it rather slowly.

6. CONCLUSION

In this paper, we considered Bayesian estimation in the presence of Gaussian noise with unknown covariance but that is independent in time. We first derived the posterior of the agent and then based on that we proposed three approximations that allow for a closed form solution for the belief update. We also presented an approach for recursive estimation of the parameter vector of interest. This result can be used in parameter estimation over networks. By computer simulations, we showed that the proposed method outperformed the least squares method in terms of mean square deviation.

7. APPENDIX

Here we show the derivation from (9) to (15). First by expanding the square terms in (9), we have

$$\begin{aligned} L_t &= \sum_{\tau=0}^{t} \boldsymbol{\theta}^\top \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{H}_\tau \boldsymbol{\theta} \\ &\quad + \sum_{\tau=1}^{t} \mathbf{y}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau + \boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1} \boldsymbol{\theta}_0 + tr(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) \\ &\quad -2\boldsymbol{\theta}^\top (\sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau + \mathbf{C}_0^{-1} \boldsymbol{\theta}_0). \end{aligned} \tag{29}$$

From (12), we have $\mathbf{C}_t^{-1}\widehat{\boldsymbol{\theta}}_t = \sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1}\mathbf{y}_\tau + \mathbf{C}_0^{-1}\boldsymbol{\theta}_0$, which implies that if we add and subtract $\widehat{\boldsymbol{\theta}}_t^\top \mathbf{C}_t^{-1}\widehat{\boldsymbol{\theta}}_t$ in the above equation, it will become

$$\begin{aligned} L_t &= (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)^\top \mathbf{C}_t^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t) - \widehat{\boldsymbol{\theta}}_t^\top \mathbf{C}_t^{-1} \widehat{\boldsymbol{\theta}}_t \\ &\quad + \sum_{\tau=1}^{t} \mathbf{y}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau + \boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1} \boldsymbol{\theta}_0 + tr(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}). \end{aligned} \tag{30}$$

With the above equation, we can write

$$\begin{aligned} S_t &= \sum_{\tau=1}^{t} \mathbf{y}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau + \boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1} \boldsymbol{\theta}_0 + tr(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) \\ &\quad - \widehat{\boldsymbol{\theta}}_t^\top \mathbf{C}_t^{-1} \widehat{\boldsymbol{\theta}}_t. \end{aligned} \tag{31}$$

Next, let $S_t^{(1)} = \sum_{\tau=1}^{t} \mathbf{y}_\tau^\top \boldsymbol{\Sigma}^{-1}\mathbf{y}_\tau - \boldsymbol{\mu}_t^\top \mathbf{M}_t^{-1}\boldsymbol{\mu}_t$ with $\mathbf{M}_t$ being defined in (15). From (14) and (15), we have

$$\boldsymbol{\mu}_t^\top \mathbf{M}_t^{-1} \boldsymbol{\mu}_t = \boldsymbol{\mu}_t^\top \sum_{\tau=1}^{t} \mathbf{H}_\tau^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_\tau, \tag{32}$$

which implies

$$S_t^{(1)} = \sum_{\tau=1}^{t} (\mathbf{y}_\tau - \mathbf{H}_\tau \boldsymbol{\mu}_\tau)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\tau - \mathbf{H}_\tau \boldsymbol{\mu}_\tau). \tag{33}$$

Similarly, we can define $S_t^{(2)} = \boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1}\boldsymbol{\theta}_0 + \boldsymbol{\mu}_t^\top \mathbf{M}_t^{-1}\boldsymbol{\mu}_t - \widehat{\boldsymbol{\theta}}_t^\top \mathbf{C}_t^{-1}\widehat{\boldsymbol{\theta}}_t$, where the last term can be expanded as

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_t^\top \mathbf{C}_t^{-1} \widehat{\boldsymbol{\theta}}_t &= (\mathbf{C}_0^{-1}\boldsymbol{\theta}_0 + \mathbf{M}_t^{-1}\boldsymbol{\mu}_t)^\top (\mathbf{C}_0^{-1} + \mathbf{M}_t^{-1})^{-1} \\ &\quad \times (\mathbf{C}_0^{-1}\boldsymbol{\theta}_0 + \mathbf{M}_t^{-1}\boldsymbol{\mu}_t) \\ &= \boldsymbol{\mu}_t^\top \mathbf{M}_t^{-1} (\mathbf{C}_0^{-1} + \mathbf{M}_t^{-1})^{-1} \mathbf{M}_t^{-1} \boldsymbol{\mu}_t \\ &\quad + \boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1} (\mathbf{C}_0^{-1} + \mathbf{M}_t^{-1})^{-1} \mathbf{C}_0^{-1} \boldsymbol{\theta}_0 \\ &\quad - 2\boldsymbol{\mu}_t^\top \mathbf{M}_t^{-1} (\mathbf{C}_0^{-1} + \mathbf{M}_t^{-1})^{-1} \mathbf{C}_0^{-1} \boldsymbol{\theta}_0. \end{aligned} \tag{34}$$

With the result in (34), one can show that $S_t^{(2)}$ can be reformulated as

$$\begin{aligned} S_t^{(2)} &= (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t)^\top \left(\mathbf{C}_0^{-1} (\mathbf{C}_0^{-1} + \mathbf{M}_t^{-1})^{-1} \mathbf{M}_t^{-1} \right) \\ &\quad \times (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t) \\ &= (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t)^\top (\mathbf{C}_0 + \mathbf{M}_t)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}_t). \end{aligned} \tag{35}$$

Noting that $S_t = S_t^{(1)} + S_t^{(2)} + tr(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1})$, by (33) and (35) we have shown that the equation (10) holds.

REFERENCES

[1] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wsns with noisy links part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[2] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1035–1048, 2010.

[3] P. M. Djurić and Y. Wang, "Distributed Bayesian learning in multiagent systems: Improving our understanding of its capabilities and limitations," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 65–76, 2012.

[4] Y. Wang and P. M. Djurić, "Sequential estimation of linear models in distributed settings," in *Proceedings of the European Signal Processing Conference*, 2013.

[5] Y. Wang and P. M. Djurić, "Reaching bayesian belief over networks in the presence of communication noise," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 591–594.

[6] L. Xiao, S. Boyd, and S.J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

[7] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[8] Z. Weng and P. M. Djurić, "Efficient estimation of linear parameters from correlated node measurements over networks," *Signal Processing Letters, IEEE (to appear)*, 2014.