

DISTRIBUTED LEAST MEAN SQUARES STRATEGIES FOR SPARSITY-AWARE ESTIMATION OVER GAUSSIAN MARKOV RANDOM FIELDS

Paolo Di Lorenzo and Sergio Barbarossa

Sapienza University of Rome, DIET, Via Eudossiana 18, 00184 Rome, Italy.

E-mail: dilorenzo,sergio@infocom.uniroma1.it

ABSTRACT

In this paper we propose distributed strategies for the estimation of sparse vectors over adaptive networks. The measurements collected at different nodes are assumed to be spatially correlated and distributed according to a Gaussian Markov random field (GMRF) model. We derive optimal sparsity-aware algorithms that incorporate prior information about the statistical dependency among observations. Simulation results show the potential advantages of the proposed strategies for online recovery of sparse vectors.

1. INTRODUCTION AND RELATED WORK

We consider the problem of distributed estimation, where a set of nodes is required to collectively estimate some vector parameter of interest from noisy measurements by relying solely on in-network processing. One typical strategy is the incremental approach [1]-[2], where each node communicates only with one neighbor at a time over a cyclic path. However, determining a cyclic path that covers all nodes is an NP-hard problem and, in addition, cyclic trajectories are prone to link and node failures. To address these difficulties, diffusion-based [3] and consensus based [4] techniques were proposed and studied in literature. In these implementations, the nodes exchange information locally and cooperate with each other without the need for a central processor. In this way, information is processed on the fly by all nodes and the data diffuse across the network by means of a real-time sharing mechanism. A characteristic of the observed signal that can be advantageously exploited to improve the estimation accuracy is the sparsity of the parameter to be estimated, as demonstrated in many recent efforts in the area of compressive sensing (CS) [5]-[6]. In particular, several algorithms for sparse adaptive filtering have been proposed based on Least-Mean Squares (LMS) [8], Recursive Least Squares (RLS) [9], and projection-based methods [10]. In subsequent studies, cooperative techniques have also been proposed for sparse adaptation over networks. Investigations on sparsity-aware, adaptive, and distributed solutions appear in [11], [12], [13], and [14]. In [11], the authors employ projection techniques onto hyperslabs and weighted ℓ_1 -balls to develop a useful sparsity-aware algorithm for distributed learning over diffusion networks. In [12], [13], the authors employed diffusion techniques that are able to identify and track sparsity over networks in a distributed manner thanks to the use of convex regularization terms. A distributed RLS algorithm with ℓ_1 -norm constraint, which enables a recursive solution, was proposed in [14].

All the previous methods considered the simple situation where the observations are uncorrelated. In some circumstances, however, this assumption is unjustified. This is the case, for example, when the

sensors monitor a field of spatially correlated values, like a temperature or atmospheric pressure field. In such cases, nearby nodes sense correlated values and then the statistical independence assumption is no longer valid. It is then of interest, in such cases, to check whether the structure of the joint probability density function (pdf) induced by the statistical dependency of the observations can be exploited to improve performance. There is indeed a broad class of observation models where the joint pdf cannot be factorized into the product of the individual pdf's pertaining to each node, but it can still be factorized into functions of subsets of variables. This is the case of Markov random fields [15]. An example of (Gaussian) Markov random field where the statistical dependencies incorporate the spatial distances was suggested in [16]. In this paper, we propose distributed LMS strategies for sparsity-aware estimation over adaptive networks, where the observations collected at different sensors are spatially correlated and distributed according to a GMRF. These solutions are able to exploit prior knowledge regarding the spatial correlation among nodes, while tracking sparsity and processing data in real-time and in a fully distributed manner. Numerical results illustrate the performance of the proposed strategies.

2. GAUSSIAN MARKOV RANDOM FIELDS

A Markov random field is represented through an undirected graph [15]. More specifically, a Markov network consists of: (a) an undirected graph $G = (V, E)$, where each vertex $v \in V$ represents a random variable and each edge represents the statistical dependency among random variables; (b) a set of potential (or compatibility) functions $\psi_c(\mathbf{x}_c)$, that associate a non-negative number to the cliques¹ of G . Let us denote by \mathcal{C} the set of all cliques present in the graph. The random vector \mathbf{x} is Markovian if its joint pdf admits the factorization

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c), \quad (1)$$

where \mathbf{x}_c denotes the vector of variables belonging to the clique c . The term Z is simply a normalization factor necessary to guarantee that $p(\mathbf{x})$ is a valid pdf. A node p is conditionally independent of another node q in the Markov network, given some set S of nodes, if every path from p to q passes through a node in S . Hence, representing a set of random variables by drawing the correspondent Markov graph is a meaningful pictorial way to identify the conditional dependencies occurring across the random variables. This is the so called *pairwise Markov property*. This property states that the joint pdf factorizes in terms that contain only variables whose vertices are neighbors. An important example of jointly Markov random variables is the Gaussian Markov Random Field (GMRF), characterized

This work has been supported by TROPIC Project, Nr. 318784.

¹A clique is a subset of nodes which are fully connected and maximal.

by having a joint pdf expressed as

$$p(\mathbf{x}) = \sqrt{\frac{|\mathbf{B}|}{(2\pi)^N}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{B}(\mathbf{x} - \boldsymbol{\mu})}, \quad (2)$$

where $\boldsymbol{\mu} = \mathbb{E}\{\mathbf{x}\}$ is the expected value of \mathbf{x} , $\mathbf{B} = \mathbf{C}^{-1}$ is the so called *precision* matrix, with $\mathbf{C} = \mathbb{E}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\}$ denoting the covariance matrix of \mathbf{x} , which is assumed to be positive definite. In this case, the *Markovianity* of \mathbf{x} manifests itself through the *sparsity* of the precision matrix, i.e., the coefficient b_{ij} of \mathbf{B} is different from zero if and only if nodes i and j are neighbors in the dependency graph. The following result from [16] provides an explicit expression between the coefficients of the covariance and the precision matrices for acyclic graphs. The elements of the precision matrix $\mathbf{B} = \{b_{ij}\}$, for a covariance matrix $\mathbf{C} = \{c_{ij}\} \succ 0$ and an acyclic dependency graph, are given by:

$$b_{ii} = \frac{1}{c_{ii}} + \frac{c_{ij}^2/c_{ii}}{c_{ii}c_{jj} - c_{ij}^2}; \quad b_{ij} = \begin{cases} \frac{-c_{ij}}{c_{ii}c_{jj} - c_{ij}^2}, & j \in \mathcal{M}_i; \\ 0, & \text{o.w.} \end{cases} \quad (3)$$

where $\mathcal{M}_i = \{j \in V : b_{ij} > 0\}$ denotes the set of neighbors of node i in the dependency graph. Let us assume that $c_{ii} = \sigma^2$, for all i , and that the amount of correlation between the neighbors (i, j) of the dependency graph is specified by an arbitrary function $g(\cdot)$, which has the Euclidean distance d_{ij} as its argument, i.e. $c_{ij} = \sigma^2 g(d_{ij})$. Furthermore, if we assume that the function $g(\cdot)$ is a monotonically non-increasing function of the distance (since amount of correlation usually decays as nodes become farther apart) and $g(0) = \nu < 1$, exploiting a result from [16], it holds $\mathbf{C} \succ 0$.

3. SPARSE DISTRIBUTED LMS ESTIMATION

Let us consider a network composed of N nodes, where the observation $x_i[k]$ collected by node i , at time k , follows the linear model

$$x_i[k] = \mathbf{u}_i^T[k] \boldsymbol{\theta} + v_i[k], \quad i = 1, \dots, N \quad (4)$$

where $\boldsymbol{\theta}$ is the M -size column vector to be estimated, $\mathbf{u}_i[k]$ is a known time-varying regression vector of size M and $\mathbf{v}[k] = [v_1[k], \dots, v_N[k]]^T$ is an observation noise vector such that $\mathbf{v}[k] \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, with \mathbf{C} denoting the noise covariance matrix. The interaction among the nodes is modeled as an undirected graph, which is described by a symmetric $N \times N$ adjacency matrix $\mathbf{A} := \{a_{ij}\}$, whose entries a_{ij} are either positive or zero, depending on whether there is a link between nodes i and j or not. To ensure that the data from an arbitrary node can eventually percolate through the entire network, we assume that the network graph is connected.

The goal of the network is to adaptively estimate the parameter $\boldsymbol{\theta}$ starting from the observations $x_i[k]$ in a sparsity-aware, distributed way, i.e. without requiring the presence of a sink nodes that collects all the measurements. Sparsity can be enforced for example by assuming a Bayesian approach incorporating some prior knowledge of $\boldsymbol{\theta}$. For example, $\boldsymbol{\theta}$ can be modeled as the outcome of a random vector having a Laplacian prior pdf, i.e.

$$p_{\Theta}(\boldsymbol{\theta}) = (\gamma/2)^M \exp(-\gamma \|\boldsymbol{\theta}\|_1) \quad (5)$$

with $\gamma > 0$, having denoted with $\|\boldsymbol{\theta}\|_1$ the ℓ_1 norm of $\boldsymbol{\theta}$. Following a Bayesian approach, the optimal estimate can be found as the vector that maximizes the log-likelihood function [17], i.e., as the solution of the following optimization problem:

$$\max_{\boldsymbol{\theta}} \mathbb{E} \{ \log [p(\mathbf{x}[k]/\boldsymbol{\theta}) p_{\Theta}(\boldsymbol{\theta})] \} \quad (6)$$

where $p(\mathbf{x}[k]/\boldsymbol{\theta})$ is the pdf of the observation vector $\mathbf{x}[k] = [x_1[k], \dots, x_N[k]]^T$ collected by all the nodes at time k , conditioned to $\boldsymbol{\theta}$. Let us assume the noise sequence $\mathbf{v}[k]$ is distributed according to a Gaussian Markov random field with zero-mean and precision matrix \mathbf{B} . Exploiting the joint pdf expression (2) and the linear observation model (4) in the formulation (6) and considering the Laplacian prior in (5), the optimal estimate $\hat{\boldsymbol{\theta}}[k]$ at time k is the vector that solves the following optimization problem:

$$\hat{\boldsymbol{\theta}}[k] = \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \mathbb{E} \{ \|\mathbf{x}[k] - \mathbf{U}[k]\boldsymbol{\theta}\|_{\mathbf{B}}^2 \} + \gamma \|\boldsymbol{\theta}\|_1 \quad (7)$$

where $\mathbf{U}[k] = [\mathbf{u}_1[k], \dots, \mathbf{u}_N^T[k]]^T$ and $\|\mathbf{y}\|_{\mathbf{X}}^2 = \mathbf{y}^T \mathbf{X} \mathbf{y}$, with \mathbf{X} denoting a generic positive definite matrix. Due to the presence of the weighted norm in (7), the problem does not seem to be amenable for a distributed solution. However, since the precision matrix \mathbf{B} reflects the structure of the dependency graph, we can write

$$V(\mathbf{x}[k]) := \frac{1}{2} \|\mathbf{x}[k] - \mathbf{U}[k]\boldsymbol{\theta}\|_{\mathbf{B}}^2 = \sum_{i=1}^N \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}) \quad (8)$$

with $\mathbf{x}_i[k] = [x_i[k], \{x_j[k]\}_{j \in \mathcal{M}_i, j > i}]^T$, and

$$\begin{aligned} \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}) &:= \frac{1}{2} b_{ii} (x_i[k] - \mathbf{u}_i^T[k]\boldsymbol{\theta})^2 \\ &+ \sum_{j \in \mathcal{M}_i, j > i} b_{ij} (x_j[k] - \mathbf{u}_j^T[k]\boldsymbol{\theta})(x_i[k] - \mathbf{u}_i^T[k]\boldsymbol{\theta}). \end{aligned} \quad (9)$$

Thus, exploiting (8) in (7), the optimization problem becomes

$$\hat{\boldsymbol{\theta}}[k] = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^N \mathbb{E} \{ \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}) \} + \gamma \|\boldsymbol{\theta}\|_1. \quad (10)$$

The problem in (10) still needs a centralized solution due to the presence of the parameter $\boldsymbol{\theta}$, which is common to every node in the network. To find a distributed solution of (10), we introduce the local estimates $\boldsymbol{\theta}_i$, for each node, and add the constraint that all the local estimates must be equal to a common, instrumental variable \mathbf{z} [18]. Thus, let us consider the convex constrained minimization problem:

$$\begin{aligned} \min_{\{\boldsymbol{\theta}_i\}_{i=1}^N} & \sum_{i=1}^N \mathbb{E} \{ \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i) \} + \gamma \|\boldsymbol{\theta}_i\|_1 \\ \text{s.t.} & \quad \boldsymbol{\theta}_i = \mathbf{z}, \quad \forall i = 1, \dots, N. \end{aligned} \quad (11)$$

To solve the problem in (11), let us consider the quadratically augmented Lagrangian function given by:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}, \mathbf{z}, \boldsymbol{\lambda}) &= \sum_{i=1}^N \mathbb{E} \{ \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i) \} + \gamma \|\mathbf{z}\|_1 \\ &+ \sum_{i=1}^N \lambda_i^T (\boldsymbol{\theta}_i - \mathbf{z}) + \frac{\rho}{2} \sum_{i=1}^N \|\boldsymbol{\theta}_i - \mathbf{z}\|^2 \end{aligned} \quad (12)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^N$, $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_i\}_{i=1}^N$ are the Lagrange multipliers associated to the equality constraints in (11), and $\rho > 0$ is a coefficient penalizing the violation of the constraints. We will now resort to the Alternating Direction Method of Multipliers (AD-MoM) [18] to iteratively minimize (12) through a set of simple recursions that update $(\boldsymbol{\Theta}, \mathbf{z}, \boldsymbol{\lambda})$ in a fully distributed fashion. The first step implements recursions of the local estimates $\boldsymbol{\Theta}$ obtained by minimizing (12) using block coordinate descent, i.e., (12) is minimized with regards to $\boldsymbol{\Theta}$

assuming the other variables \mathbf{z} and $\boldsymbol{\lambda}$ are fixed. The separable structure of (11) is inherited by the augmented Lagrangian, and therefore $\Theta[k+1] = \arg \min_{\Theta} \mathcal{L}(\Theta, \mathbf{z}[k], \boldsymbol{\lambda}[k])$ decouples into N simpler minimization subproblems:

$$\begin{aligned} \boldsymbol{\theta}_i[k+1] = \arg \min_{\boldsymbol{\theta}_i} & \left[\mathbb{E} \{ \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i) \right. \\ & \left. + \boldsymbol{\lambda}_i^T[k](\boldsymbol{\theta}_i - \mathbf{z}[k]) + \frac{\rho}{2} \|\boldsymbol{\theta}_i - \mathbf{z}[k]\|^2 \right]. \end{aligned} \quad (13)$$

It is important to notice that, in this case, even if the global problem concerning the minimization of the augmented Lagrangian in (12) is certainly convex, the local problem in (13) is not necessarily convex because there is no guarantee that the term $\phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i)$ is a positive definite function. Nevertheless, the quadratic penalty in (13) can make every local problem in (13) convex, provided that the coefficient ρ is chosen sufficiently large. At the same time, at convergence, the penalty goes to zero by virtue of the constraints in (11), thus not inducing an undesired bias on the final result. Then, since at large ρ the cost function in (13) is convex and differentiable, the first-order necessary condition is also sufficient for optimality, thus getting

$$\mathbb{E} \{ \nabla_{\boldsymbol{\theta}_i} \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i) \} + \boldsymbol{\lambda}_i[k] + \rho(\boldsymbol{\theta}_i - \mathbf{z}[k]) = \mathbf{0}. \quad (14)$$

The estimate update can be obtained as the root of the equation in (14). Nevertheless, since the local covariance information in $\mathbb{E} \{ \nabla_{\boldsymbol{\theta}_i} \phi_i(\mathbf{x}_i[k]; \boldsymbol{\theta}_i) \}$ (see eq. (9)) are not known, the root of (14) is not computable in closed form. Exploiting (9) and motivated by stochastic approximation techniques [19], which find the root of an unknown function given a time-series of noisy observations, the proposed recursion is given by

$$\begin{aligned} \boldsymbol{\theta}_i[k+1] = \boldsymbol{\theta}_i[k] + \mu_i & \left[b_{ii} \mathbf{u}_i[k] \left(x_i[k] - \mathbf{u}_i^T[k] \boldsymbol{\theta}_i[k] \right) \right. \\ & + \sum_{j \in \mathcal{M}_i, j > i} b_{ij} \left[(x_j[k] \mathbf{u}_i[k] + x_i[k] \mathbf{u}_j[k]) - \left(\mathbf{u}_j[k] \mathbf{u}_i^T[k] \right. \right. \\ & \left. \left. + \mathbf{u}_i[k] \mathbf{u}_j^T[k] \right) \boldsymbol{\theta}_i[k] \right] - \boldsymbol{\lambda}_i[k] - \rho(\boldsymbol{\theta}_i[k] - \mathbf{z}[k]) \right], \end{aligned} \quad (15)$$

where $\mu_i > 0$ is a constant step-size. In order to compute (15), we assume each node i knows the coefficients $\{b_{ij}\}$ of the i -th row of the precision matrix \mathbf{B} . The second step entails updating the global variable $\mathbf{z}[k]$. The recursions are then obtained by minimizing (12) keeping fixed the variables $\Theta[k+1]$ and $\boldsymbol{\lambda}[k]$, thus leading to the closed form solution:

$$\mathbf{z}[k+1] = \frac{1}{\rho N} T_{\gamma} \left(\sum_{i=1}^N \boldsymbol{\lambda}_i[k] + \rho \sum_{i=1}^N \boldsymbol{\theta}_i[k+1] \right), \quad (16)$$

where $T_{\gamma}(x)$ is a threshold function defined over each component as

$$T_{\gamma}(x) = \begin{cases} x - \gamma, & x > \gamma; \\ 0, & -\gamma \leq x \leq \gamma; \\ x + \gamma, & x < -\gamma. \end{cases} \quad (17)$$

The third step is a dual variable update, whose goal is to maximize the dual function, as in the dual ascent method. The pertinent recursions for the multipliers' update are then

$$\boldsymbol{\lambda}_i[k+1] = \boldsymbol{\lambda}_i[k] + \rho(\boldsymbol{\theta}_i[k+1] - \mathbf{z}[k+1]) \quad (18)$$

for $i = 1, \dots, N$. Recursions (15), (16), and (18) constitute the proposed algorithm, which can be arbitrarily initialized. From this formulation, we can see that the first and third steps can run in parallel,

over each node. The only step that requires an exchange of values among the nodes is the second step that requires the computation of an average value, which can be computed in a distributed fashion through the use of a consensus algorithm [20] over a connected graph. Let us introduce the notation \bar{x} as the averaging operation across the nodes, i.e. $\bar{x} := \frac{1}{N} \sum_{i=1}^N x_i$. The algorithm is summarized in Table 1; we will refer to it as the Sparse Distributed LMS algorithm over GMRFs (SD-LMS-GMRF).

Table 1: SD-LMS-GMRF

Each node initializes $\boldsymbol{\theta}_i[0]$, $\mathbf{z}[0]$, and $\boldsymbol{\lambda}_i[0]$ randomly. Then, it repeats the following steps for $k \geq 0$:

- 1) $\boldsymbol{\theta}_i[k+1] = \boldsymbol{\theta}_i[k] + \mu_i \left[b_{ii} \mathbf{u}_i[k] \left(x_i[k] - \mathbf{u}_i^T[k] \boldsymbol{\theta}_i[k] \right) \right. \\ \left. + \sum_{j \in \mathcal{M}_i, j > i} b_{ij} \left[(x_j[k] \mathbf{u}_i[k] + x_i[k] \mathbf{u}_j[k]) - \left(\mathbf{u}_j[k] \mathbf{u}_i^T[k] \right. \right. \right. \\ \left. \left. \left. + \mathbf{u}_i[k] \mathbf{u}_j^T[k] \right) \boldsymbol{\theta}_i[k] \right] - \boldsymbol{\lambda}_i[k] - \rho(\boldsymbol{\theta}_i[k] - \mathbf{z}[k]) \right];$
- 2) Run a consensus loop to compute $\overline{\boldsymbol{\lambda}[k]}$ and $\overline{\boldsymbol{\theta}[k+1]}$;
- 3) $\mathbf{z}[k+1] = \frac{1}{\rho N} T_{\gamma} \left(N \overline{\boldsymbol{\lambda}[k]} + \rho N \overline{\boldsymbol{\theta}[k+1]} \right);$
- 4) $\boldsymbol{\lambda}_i[k+1] = \boldsymbol{\lambda}_i[k] + \rho(\boldsymbol{\theta}_i[k+1] - \mathbf{z}[k+1]);$

As we can see from Table 1, the SD-LMS-GMRF algorithm has inherently two time-scales, i.e., the sensing time k , and one consensus loop per iteration k , which is needed to fuse the data coming from all the nodes in the network. In principle, one could argue that running many consensus iterations may not be a problem in a stationary environment. However, when the WSN is required to track a time-varying parameter vector, one cannot afford significant delays in-between consecutive sensing instants. One possible way to overcome this hurdle is to run a single consensus iteration per acquired observation $(x_i[k], \mathbf{u}_i[k])$, which can be simply written as a local weighted average among neighbors, i.e.,

$$\overline{\boldsymbol{\theta}_i[k+1]} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{\theta}_j[k+1], \quad (19)$$

where \mathcal{N}_i is the spatial neighborhood of node i comprised itself, and $\mathbf{W} = \{w_{ij}\} \in \mathbb{R}^{N \times N}$ are real, non-negative coefficients satisfying $\mathbf{W}\mathbf{1} = \mathbf{1}$. The combination of intermediate estimates in (19) reminds the diffusion approach followed in [3], where several rules were proposed to select the coefficients w_{ij} in order to improve the estimation and tracking performance of the resulting algorithms. Similar arguments follow to compute $\overline{\boldsymbol{\lambda}_i[k]}$. Thus, exploiting the combination step in (19) to evaluate $\overline{\boldsymbol{\theta}_i[k+1]}$ and $\overline{\boldsymbol{\lambda}_i[k]}$ at step 2 in Table 1, one arrives at a Single Time Scale Sparse Distributed LMS algorithm over GMRF (STS-SD-LMS-GMRF), which is more suitable for operating in nonstationary environments.

Up to now, we have considered the ℓ_1 norm constraint in (7), which is commonly known as the Lasso [7]. We denote the strategy using the function in (17) as the ℓ_1 -SD-LMS-GMRF algorithm. It is well known that the Lasso leads to the presence of a strong bias in the case of estimation of less sparse vectors. To reduce the effect of the bias, we consider the non-negative Garotte estimator as in [21],

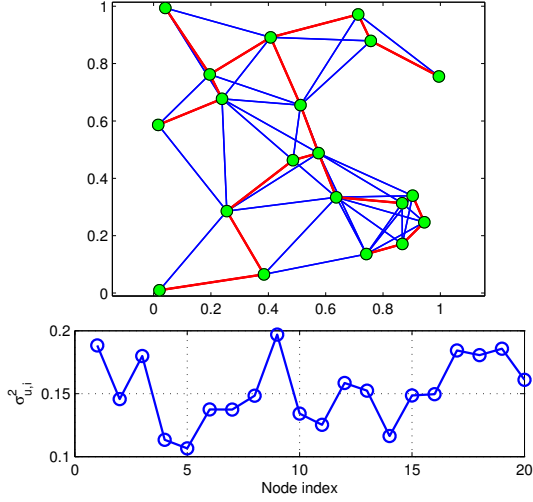


Fig. 1: Adjacency (Blue line) and Dependency (Red line) graphs (top). Regressor variances (bottom).

whose thresholding function is defined as a vector whose entries are derived applying the threshold

$$T_\gamma(x) = \begin{cases} x(1 - \gamma^2/x^2), & |x| > \gamma; \\ 0, & -\gamma \leq x \leq \gamma; \end{cases} \quad (20)$$

We denote the algorithm using (20) as the *G-SD-LMS-GMRF*.

Numerical Results : We consider a connected network composed of 20 nodes. The topology of the network is shown in Fig. 1, where the blue links represent the communication graph, whereas the red links depict the statistical dependency graph. Each red link is also supported by a communication channel so that the dependency graph can be seen as a sub-graph of the communication graph. The regressors $\mathbf{u}_i[k]$ have size $M = 50$ and are zero-mean white Gaussian distributed with covariance matrices $R_{u_i} = \sigma_{u_i}^2 I_M$, with $\sigma_{u_i}^2$ shown in Fig. 1 (bottom). Since the dependency graph in Fig. 1 is acyclic, we compute the precision matrix as in (3) with $c_{ii} = \sigma^2 = 0.0157$ and $c_{ij} = \sigma^2 \nu \exp(-\kappa \cdot d_{ij})$, where d_{ij} is the Euclidean distance among nodes i and j , $\nu < 1$ is the nugget parameter, and $\kappa \geq 0$ is a correlation coefficient. To illustrate the steady-state performance of the proposed strategies, in Fig. 2, we report the steady-state network Mean Square Deviation, i.e., $\text{MSD} = \lim_{k \rightarrow \infty} \mathbb{E} \sum_{i=1}^N \|\mathbf{z}_i[k] - \boldsymbol{\theta}\|^2 / N$ versus the number of non-zero components (set equal to one) of the vector parameter $\boldsymbol{\theta}$, for 3 different algorithms: the sparsity agnostic STS-SD-LMS-GMRF with $\gamma = 0$, the ℓ_1 -STS-SD-LMS-GMRF with $\gamma = 3$, and the G-STS-SD-LMS-GMRF with $\gamma = 45$, which are described in Table 1, with combination steps as in (19) and thresholding functions in (17) and (20), respectively. The results are averaged over 100 independent realizations. The combination matrix \mathbf{W} is chosen such that $w_{ij} = 1/|\mathcal{N}_i|$ for all i . The other parameters are chosen as $\nu = 0.7$, $\kappa = 0.5$, $\mu_i = 10^{-3} \forall i$, and $\rho = 500$. As we can notice from Fig. 2, when the vector is very sparse both sparsity-aware strategies yield better steady-state performance than the sparsity agnostic algorithm. When the vector $\boldsymbol{\theta}$ is less sparse, the ℓ_1 -STS-SD-LMS-GMRF performs worse than the sparsity-agnostic counterpart due to the dominant effect of the bias introduced by (17). The G-STS-SD-LMS-GMRF greatly outperforms both ℓ_1 -STS-SD-LMS-GMRF and sparsity-agnostic algorithms, while matching the performance of the latter only when the vector is completely non sparse.

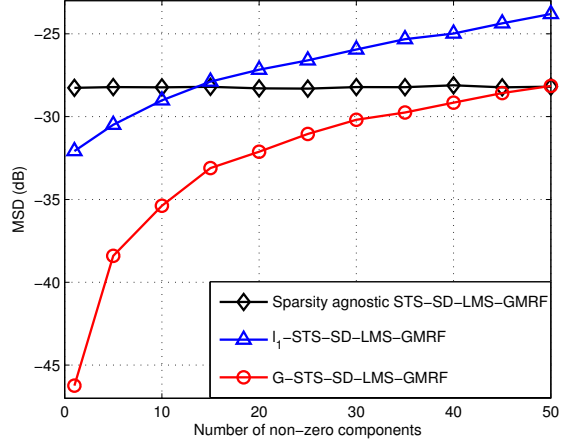


Fig. 2: Network MSD versus number of non-zero components of the true vector parameter, considering different algorithms.

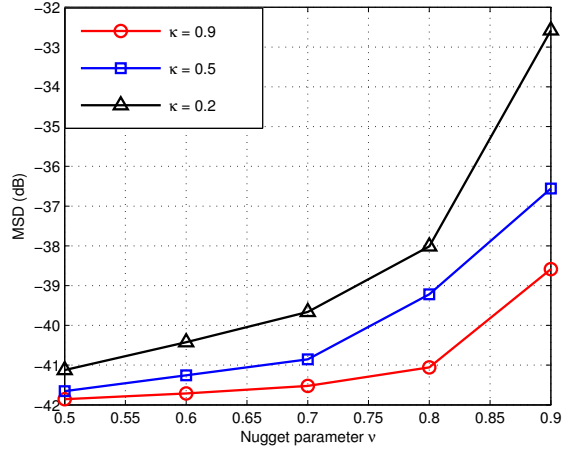


Fig. 3: Network MSD versus nugget parameter, considering different values of the correlation coefficient κ .

To assess the sensitivity of the proposed strategies to variations in the parameters describing the GMRF, in Fig. 3, we report the steady-state network MSD of the G-STS-SD-LMS-GMRF versus the nugget parameter ν , considering different values of the correlation coefficient κ . The results are averaged over 100 independent realizations. We consider a vector parameter $\boldsymbol{\theta}$ with only 5 elements set equal to one, which have been randomly chosen. The other parameters are chosen as $\mu_i = 6 \times 10^{-4} \forall i$, $\rho = 500$, and $\gamma = 45$. As we can see from Fig. 3, the performance of the algorithm gets worse by increasing the correlation among the nodes observations.

4. CONCLUSIONS

In this paper we have proposed adaptive strategies for distributed estimation of sparse vectors, which are able to exploit the underlying structure of the statistical dependency among observations collected by network nodes at different spatial locations. Two different thresholding functions, namely, the Lasso and the Garotte estimators, have been exploited to improve performance under sparsity. Simulation results illustrate the potential advantages achieved by these strategies for online, distributed, sparse data recovery.

5. REFERENCES

- [1] D. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.
- [2] C. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4064–4077, 2007.
- [3] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Transactions on Signal Processing*, vol. 58, pp. 1035–1048, March 2010.
- [4] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2365–2382, Jun. 2009.
- [5] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [6] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, March 2007.
- [7] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996.
- [8] Y. Chen, Y. Gu, and A.O. Hero, "Sparse LMS for system identification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3125–3128, Taipei, May 2009.
- [9] D. Angelosante, J.A. Bazerque, and G.B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the ℓ_1 -norm," *IEEE Trans. on Signal Processing*, vol. 58, no. 7, pp. 3436–3447, July, 2010.
- [10] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted ℓ_1 balls," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 936–952, March, 2010.
- [11] S. Chouvardas, K. Slavakis, Y. Kopsinis, S. Theodoridis, "A sparsity-promoting adaptive algorithm for distributed learning," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5412–5425, Oct. 2012.
- [12] Y. Liu, C. Li and Z. Zhang, "Diffusion sparse least-mean squares over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4480–4485, Aug. 2012.
- [13] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Transactions on Signal Processing*, vol. 61, no. 6, pp. 1419–1433, 15 March 2013.
- [14] S. Sardellitti and S. Barbarossa, "Distributed RLS estimation for cooperative sensing in small cell networks," *IEEE Inter. Conf. on Acoustics, Speech and Signal Process.*, Vancouver, Canada, 2013.
- [15] S. L. Lauritzen, *Graphical Models*, Oxford University Press, 1996.
- [16] A. Anandkumar, L. Tong, A. Swami, "Detection of Gauss-Markov Random Fields With Nearest-Neighbor Dependency," *IEEE Trans. on Information Theory*, Vol. 55, pp. 816–827, Feb. 2009.
- [17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.
- [18] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*, Belmont, MA, Athena Scientific, 1997.
- [19] M. Nevelson and R. Hasminskii, "Stochastic approximation and recursive estimation," Providence, Rhode Island: American Mathematical Society, 1973.
- [20] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.
- [21] M. Yuan, Y. Lin, "On the non-negative garrotte estimator," *J. Royal Statist. Soc.*, vol. 69, pp. 143161, 2007.