Diffusion Estimation Over Cooperative Multi-Agent Networks With Missing Data

Diffusion Estimation Over Cooperative Multi-Agent Networks With Missing Data

Mohammad Reza Gholami, *Member, IEEE*, Magnus Jansson, *Member, IEEE*,
Erik G. Ström, *Senior Member, IEEE*, and Ali H. Sayed, *Fellow, IEEE*

***Abstract*—In many fields, and especially in the medical and social sciences and in recommender systems, data are gathered through clinical studies or targeted surveys. Participants are generally reluctant to respond to all questions in a survey or they may lack information to respond adequately to some questions. The data collected from these studies tend to lead to linear regression models where the regression vectors are only known partially: some of their entries are either missing completely or replaced randomly by noisy values. In this work, assuming missing positions are replaced by noisy values, we examine how a connected network of agents, with each one of them subjected to a stream of data with incomplete regression information, can cooperate with each other through local interactions to estimate the underlying model parameters in the presence of missing data. We explain how to adjust the distributed diffusion through (de)regularization in order to eliminate the bias introduced by the incomplete model. We also propose a technique to recursively estimate the (de)regularization parameter and examine the performance of the resulting strategy. We illustrate the results by considering two applications: one dealing with a mental health survey and the other dealing with a household consumption survey.**

***Index Terms*– Missing data, linear regression, mean-square-error, regularization, distributed estimation, diffusion strategy.**

I. Introduction

In data gathering procedures through clinical studies or targeted surveys, it is common that some components of the data are missing or left unobserved. For example, in a clinical study, a participant may be reluctant to answer some questions or may drop out of the survey and never return it [2]. Likewise, in a recommender system using content based filtering [3], a participant may prefer to leave some questions unanswered. The phenomenon of missing data is prevalent in many fields including the social sciences, medical sciences, econometrics, survival data analysis, and machine learning [2], [4]–[11]. In general, three patterns are considered for missing data: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) [12]–[14]. In MAR patterns, the probability of missing a variable in data gathering is related not only to the value of the variable but also to some other measured variables [2]. In comparison, in MCAR scenarios, the probability of missing a variable is solely related to the value of that variable [12]. For example, if an observation is below a threshold, it may not be observed. In MNAR, the probability of missing depends on unobserved values. For example, in a cancer trial some patients may be feeling weak and unable to participate in the study [2]. In general it may not be easy to verify which pattern of missing is happening for a study. For a survey on different missing patterns, the reader may refer to, e.g., [2, Ch. 1] and the references therein.

There have been many studies on inference problems for missing data with several useful techniques proposed to address the challenges associated with censored data. Many of the approaches are based on heuristic methods, as already noted in [15]. There are broadly two classes of techniques to deal with estimation in the presence of missing data: imputation and deletion (listwise or pairwise) [2], [10], [16]–[20]. If the positions of the missing data are known in advance, then they can either be replaced by some deterministic or random values (sometimes called single or multiple imputation), or the corresponding data can be removed from the dataset altogether. Removing data generally leads to performance degradation, although this practice is accepted in some statistical software packages, e.g., SAS, SPSS, GAUSS, and CDAS, mainly because of ease of implementation. In contrast, data imputation results in biased estimates [6], [14], [16], [19]. For a discussion on different types of deletion and imputation in missing data analysis, see [2, Ch. 2].

When data imputation is assumed, one may resort to a maximum likelihood procedure to estimate the missing data if the distribution of the data happens to be known beforehand [21], [22]. For example, the expectation-maximization (EM) technique provides one useful solution method [9]. However, the EM implementation tends to be computationally intensive and can suffer from convergence issues [23]. If the positions of the missing data are unknown, a mixture model may be used to describe the data model [24] and the EM algorithm can be subsequently applied to estimate the missing data [25], [26]. However, the number of components can become large in high dimensions. For other possible recovery techniques including ordinary least-squares, inverse probability weighting, and fully

M. R. Gholami is with Campanja AB, Stockholm SE-111 57, Sweden (e-mail: reza@campanja.com)

M. Jansson is with the ACCESS Linnaeus Center, Electrical Engineering, KTH–Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: janssonm@kth.se)

E. G. Ström is with the Division of Communication Systems, Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden (e-mail:erik.strom@chalmers.se).

A. H. Sayed is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 (e-mail: sayed@ee.ucla.edu).

This work was initiated while M. R. Gholami was a visiting student at the UCLA Adaptive Systems Laboratory. The work was supported in part by the Swedish Research Council (contract no. 2007-6363) and the Access center and in part by Ericsson's Research Foundation (Reference: FOSTIFT-12:021 and FOSTIFT-13:031). The work of A. H. Sayed was supported in part by NSF grants CCF-1524250 and ECCS-1407712. A short and preliminary version of this work was presented in the conference publication [1].

Bayseian methods, the reader may refer to, e.g., [4], [27]–[30].

In this work, we examine how a connected network of agents, with each one of them subjected to a stream of data with incomplete regression information, can cooperate with each other to estimate the underlying model parameters in the presence of missing data. In particular, we assume that some entries of the regressor can be missed according to the MCAR pattern, where the missing data are replaced by random entries. Note that for nonstreaming data, there are various techniques in the literature [31]–[33] to handle an inference problem. These techniques are useful if agents are able to collect the data and save them in buffers for batch or centralized processing.

Indeed, one of the main approaches used in the literature to address estimation in the presence of missing data relies on centralized solutions, where the data is collected centrally and processed by a fusion center. This approach has the disadvantage of requiring a large amount of energy and communication and may limit scalability and robustness of the network. There are many situations, however, where the data is already available in a distributed manner (e.g., separate clinics collect surveys from their patients independently of each other). Moreover, in many of these cases, privacy and secrecy considerations may prevent the clinics or survey-collectors to share or transmit the data to a central location. In these situations, it is preferable to process the data in a distributed manner. In this approach, only estimates of the unknown parameter model, and not the raw collected data, are exchanged between neighboring nodes. This paper focuses on this approach and develops a distributed strategy for handling missing data from results collected at dispersed locations. One of the key challenges is to infuse the distributed procedure with a mechanism to combat the bias that results from the presence of censored data.

For this purpose, we shall rely on the idea of *diffusion* adaptation [34]–[38] to design a fully distributed algorithm that is able to adapt and learn from streaming data. Useful consensus strategies can also be employed [31], [39]–[43]. We focus on diffusion strategies in this work due to their proven enhanced stability ranges and improved mean-square-performance properties over consensus networks when constant step-sizes are employed for continuous adaptation and learning [35]–[37], [44]; consensus networks are sensitive to the network topology and their state error vectors can grow unbounded when constant step-sizes are used. We will explain how to adjust the diffusion strategy through (de)regularization in order to eliminate the bias introduced by imputation. We will examine the stability and performance of the proposed algorithm and provide computer simulations on two applications: one dealing with a mental health survey and the other dealing with a household consumption survey.

Notation. We use lowercase letters to denote vectors, uppercase letters for matrices, plain letters for deterministic variables, boldface letters for random variables. We use $\odot$ and $\otimes$ for the Hadamard (elementwise) and Kronecker products, respectively. In addition, $\mathrm{diag}\{x_1,\ldots,x_N\}$ denotes a diagonal matrix with diagonal elements $x_1,\ldots,x_N$. We use $\mathrm{col}\{a,b\}$ to represent a column vector with entries a and b, while I_M and 0_M denote the $M\times M$ identity and $M\times M$ zero matrices. We use $\lambda_i(A)$ to denote the $i-$th eigenvalue of matrix A. The $i-$th and $j-$th entry of a matrix A is represented by $A(i,j)$.

II. Problem statement

Consider a connected network with N agents. Each agent senses a stream of wide-sense stationary data $\{\boldsymbol{d}_k(i),\boldsymbol{u}_{k,i}\}$ that satisfy the linear regression model:

$$\boldsymbol{d}_k(i)=\boldsymbol{u}_{k,i}w^o+\boldsymbol{v}_k(i),\quad k=1,\ldots,N \tag{1}$$

where k is the node index and i is the time index. The row vector $\boldsymbol{u}_{k,i}$ denotes a zero-mean random process with covariance matrix $R_{u,k}=\mathbb{E}\,\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i}>0$, while $\boldsymbol{v}_k(i)$ is a zero-mean white noise process with variance $\sigma_{v,k}^2$. The column vector $w^o\in\mathbb{R}^M$ is the unknown parameter that the network is interested in estimating.

Remark 1: Sometimes, as illustrated by the examples discussed later in Sec. V, we may encounter data models of the form

$$\bar{\boldsymbol{d}}_k(i)=\alpha+\bar{\boldsymbol{u}}_{k,i}w^o+\boldsymbol{v}_k(i),\quad k=1,\ldots,N \tag{2}$$

where α is some deterministic intercept value, and $\bar{\boldsymbol{u}}_{k,i}$ is a row vector with nonzero mean. This model can still be reduced to (1) by subtracting the mean of the measurement from both sides of (2).

Assumption 1: Continuing with model (1), we assume that the regression and noise processes are each spatially independent and temporally white. In addition, we assume that $\boldsymbol{u}_{\ell,i}$ and $\boldsymbol{v}_k(j)$ are independent of each other for all ℓ,i,k, and j. ■

In this study, we examine the situation in which some entries in the regression vectors may be missing completely at random due to a variety of reasons, including incomplete information or censoring. We denote the incomplete regressor by $\bar{\boldsymbol{u}}_{k,i}$ and express it in the form:

$$\bar{\boldsymbol{u}}_{k,i}=\boldsymbol{u}_{k,i}(I_M-\boldsymbol{F}_{k,i})+\boldsymbol{\xi}_{k,i}\boldsymbol{F}_{k,i} \tag{3}$$

where $\boldsymbol{F}_{k,i}=\mathrm{diag}\{\boldsymbol{f}_{k,i}^1,\ldots,\boldsymbol{f}_{k,i}^M\}$ consists of random indicator variables, $\boldsymbol{f}_{k,i}^j\in\{0,1\}$. Each variable $\boldsymbol{f}_{k,i}^j$ is equal to one with some probability $p<1$ and equal to zero with probability $1-p$. The value of p represents the likelihood that the $j-$th entry of the regression vector $\boldsymbol{u}_{k,i}$ is missing at time i. In that case, the missing entry is assumed to be replaced by an entry from the zero-mean perturbation (row) variable $\boldsymbol{\xi}_{k,i}$.

The model considered in (3) encompasses different types of censoring such as *left censoring*, *right censoring*, *interval censoring*, and *random censoring* [45]–[47].[1] In the missing data literature, the position of missing data is often assumed to be known in advance [16]. In general, both scenarios with known and unknown missing positions may happen in practice. For example, a participant in a survey may leave a question unanswered (known missing position), or may report a wrong value (unknown missing position since the true value is intentionally replaced by a wrong value); the latter may be

[1] In econometrics, the use of the "*coded*" terminology (such as top-coded and bottom-coded) is more common than censoring [6], [48].

considered an outlier although usually an outlier is defined as a measurement that is distant from other measurements. For a known missing position, the imputation technique can then be used to fill the position of data left unobserved yielding a modified data model. If the analyst for the modified model is not aware of the imputed position, we can still consider model (3) with the assumption that the imputer and the analyst are different and do not exchange information about the missing position.

Remark 2: Model (3) is sometimes referred to as a mixture model in the literature [26], [49]–[51]. In such models, it is mainly assumed that the components follow Gaussian distributions with different means or variances (or both). However, in this study, we have no particular assumptions about the distribution of the components or their first and second-order moments. For mixture model approaches, if the distribution of the data deviates from the normal distribution, a large number of Gaussian components is then mixed to model the missing data [52], [53]. In these cases, the design of the algorithms can become more challenging.

Assumption 2: We assume that the random variables $\boldsymbol{u}_{k,i}, \boldsymbol{f}_{k,i}^{j}$, and $\boldsymbol{\xi}_{k,i}$ are independent of each other. We also assume that the random process $\boldsymbol{\xi}_{k,i}$ is temporally white and spatially independent with covariance matrix $\mathbb{E}\,\boldsymbol{\xi}_{k,i}^{*}\boldsymbol{\xi}_{k,i} = \sigma_{\xi,k}^{2} I_M$. ■

From model (1), the minimum mean-square-error (MMSE) estimate of the unknown vector w^o based on the data collected at node k is given by [54]:

$$w_k^o = R_{u,k}^{-1}\, r_{du,k} \tag{4}$$

where

$$r_{du,k} \triangleq \mathbb{E}\,\boldsymbol{d}_k(i)\boldsymbol{u}_{k,i}^{*}. \tag{5}$$

It is easy to verify from (1) that $w_k^o = w^o$ so that the MMSE solution allows node k to recover the unknown w^o if the actual moments $\{R_{u,k}, r_{du,k}\}$ happen to be known. The resulting mean-square-error is [54]:

$$\begin{aligned} J_{k,\min} \triangleq J_k(w_k^o) &= \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w_k^o|^2 \\ &= \sigma_{d,k}^2 - r_{du,k}^{*} R_{u,k}^{-1} r_{du,k} \\ &= \sigma_{v,k}^2. \end{aligned} \tag{6}$$

Let us now investigate the estimate that would result if we use instead the censored regressor, $\bar{\boldsymbol{u}}_{k,i}$. The estimate in this case is given by

$$\bar{w}_k^o = R_{\bar{u},k}^{-1}\, r_{d\bar{u},k} \tag{7}$$

with the covariance matrix $R_{\bar{u},k}$ computed as

$$\begin{aligned} R_{\bar{u},k} &\triangleq \mathbb{E}\bar{\boldsymbol{u}}_{k,i}^{*}\bar{\boldsymbol{u}}_{k,i} \\ &\overset{(a)}{=} \mathbb{E}\{(I - \boldsymbol{F}_{k,i}) R_{u,k} (I - \boldsymbol{F}_{k,i})\} + \sigma_{\xi,k}^2 \mathbb{E}\{\boldsymbol{F}_{k,i}^2\} \\ &= R_{u,k} - P_1 \odot R_{u,k} + p\sigma_{\xi,k}^2 I_M \\ &\triangleq R_{u,k} + R_{r,k} \end{aligned} \tag{8}$$

where in (a) we used the independence of $\boldsymbol{u}_{k,i}$ and $\boldsymbol{\xi}_{k,i}$ (Assumption 2) and where the matrix $R_{r,k}$ is defined as

$$R_{r,k} \triangleq -P_1 \odot R_{u,k} + p\sigma_{\xi,k}^2 I_M \tag{9}$$

in terms of

$$P_1 \;\triangleq\; (2p - p^2)\mathbb{1}_M \mathbb{1}_M^T - (p - p^2) I_M. \tag{10}$$

Here, we are using the notation $\mathbb{1}_M$ to denote the M-column vector with all its entries equal to one. Likewise, the cross correlation vector $r_{d\bar{u},k}$ in (7) is given by

$$\begin{aligned} r_{d\bar{u},k} &\triangleq \mathbb{E}\,\boldsymbol{d}_k(i)\bar{\boldsymbol{u}}_{k,i}^{*} \\ &= \mathbb{E}\boldsymbol{d}_k(i)(I - \boldsymbol{F}_{k,i})\boldsymbol{u}_{k,i}^{*} + \underbrace{\mathbb{E}\boldsymbol{d}_k(i)\boldsymbol{F}_{k,i}\boldsymbol{\xi}_{k,i}^{*}}_{=0} \\ &= (1-p) r_{du,k}. \end{aligned} \tag{11}$$

We assume that the perturbed matrix $R_{\bar{u},k}$ remains invertible. Applying the matrix inversion lemma [55] to (8), we can relate $\bar{w}_k^o$ from (7) to w_k^o from (4) as follows:

$$\bar{w}_k^o = (I_M - Q_k)(1-p) w^o \tag{12}$$

in terms of the matrix

$$Q_k \triangleq R_{u,k}^{-1} R_{r,k} (I_M + R_{u,k}^{-1} R_{r,k})^{-1}. \tag{13}$$

It is observed from (12) that the new estimate is biased relative to w^o. It is also obvious that the mean-square-error that results from using (7) is lower bounded by $J_{k,\min}$ from (6), i.e.,

$$J_{\mathrm{miss},k} \triangleq J_k(\bar{w}_k^o) \geq J_{k,\min}. \tag{14}$$

To mitigate the bias problem, we extend a construction used in [56] in a different context and associate an alternative (de-)regularized quadratic cost with each agent k, defined as follows:

$$J_k'(w) \triangleq \mathbb{E}|\boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i} w|^2 - \|w\|_{T_k}^2 \tag{15}$$

where T_k is a symmetric matrix to be chosen. The stationary point of (15) is easily seen to occur at the following location, where we again assume that the coefficient matrix $(R_{\bar{u},k} - T_k)$ is invertible as needed:

$$\bar{w}_k^o = (1-p)(R_{\bar{u},k} - T_k)^{-1} r_{du,k}. \tag{16}$$

Therefore, if we refer to (8) and (9), we observe that if we select T_k as

$$\begin{aligned} T_k &= pR_{u,k} + R_{r,k} \\ &= (p - p^2) I_M \odot R_{u,k} + p\sigma_{\xi,k}^2 I_M - (p - p^2) R_{u,k} \end{aligned} \tag{17}$$

then the solution $\bar{w}_k^o$ from (16) will agree with the unbiased original estimate w_k^o from (4), i.e., $\bar{w}_k^o = w_k^o$.

Now note that since $p - p^2$ is nonnegative for $0 \leq p < 1$, the matrix T_k in (17) is seen to be the difference of two nonnegative definite matrices. Therefore, T_k is in general *sign* indefinite. More importantly, if the de-regularization matrix T_k is selected as in (17), the cost function in (15) becomes *strongly* convex with a unique minimizer. This is because the Hessian matrix of $J_k'(w)$ relative to w is positive-definite since

$$\nabla_w^2 J_k'(w) \;=\; 2(R_{\bar{u},k} - T_k) = 2(1-p) R_{u,k} > 0 \tag{18}$$

for $0 \leq p < 1$.

III. Adaptive Distributed Strategy

In this section, we develop a distributed strategy to recover w^o from missing data by relying on the modified local cost (15). For the remainder of the paper, we represent the covariance matrix of the regression data in the following form.

Assumption 3: The covariance matrix $R_{u,k}$ is diagonal, which is satisfied if the entries of the regression vector $\boldsymbol{u}_{k,i}$ are uncorrelated with each other. ■

Under Assumption 3, it holds that

$$R_{r,k} = -pR_{u,k} + p\sigma_{\xi,k}^2 I_M \quad (19)$$

$$R_{\bar{u},k} = (1-p)R_{u,k} + p\sigma_{\xi,k}^2 I_M \quad (20)$$

$$T_k = p\sigma_{\xi,k}^2 I_M \quad (21)$$

where T_k is now nonnegative-definite. We shall first assume that the parameters p and $\sigma_{\xi,k}^2$ are known. Later, we estimate $\sigma_{\xi,k}^2$ from the data, assuming an estimate for p is available (a condition that is generally satisfied in practice — see Sec. V-A where experimental settings are described).

To develop a distributed algorithm, we let $\mathcal{N}_k$ denote the set of neighbors of agent k. The network then seeks to solve:

$$\min_{w\in\mathbb{R}^M} \sum_{k=1}^{N} J_k'(w). \quad (22)$$

Following arguments similar to [36], [37], we can derive the following modified Adapt-then-Combine (mATC) diffusion strategy for the case of missing data:

$$\begin{cases} \boldsymbol{e}_k(i) = \boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}\boldsymbol{w}_{k,i-1} \\ \boldsymbol{\phi}_{k,i} = (1+\mu_k p\sigma_{\xi,k}^2)\boldsymbol{w}_{k,i-1} + \mu_k \bar{\boldsymbol{u}}_{k,i}^* \boldsymbol{e}_k(i) \\ \boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \end{cases} \quad (23)$$

where μ_k is a small step-size parameter and the convex combination coefficients $\{a_{\ell k}\}$ satisfy [36], [37], [57]:

$$a_{\ell k} \geq 0, \quad \sum_{\ell\in\mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k. \quad (24)$$

A. Estimation of Regularization Parameter

The distributed algorithm (23) requires knowledge of the censoring noise variance, $\sigma_{\xi,k}^2$. We now suggest one way to estimate this noise power. From (3), we write

$$\boldsymbol{u}_{k,i} = \bar{\boldsymbol{u}}_{k,i} + \boldsymbol{u}_{k,i}\boldsymbol{F}_{k,i} - \boldsymbol{\xi}_{k,i}\boldsymbol{F}_{k,i}, \quad (25)$$

and use this relation to re-write the measurement model (1) in terms of $\bar{\boldsymbol{u}}_{k,i}$ as follows:

$$\boldsymbol{d}_k(i) = \bar{\boldsymbol{u}}_{k,i}w^o + (\boldsymbol{u}_{k,i} - \boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}w^o + \boldsymbol{v}_k(i). \quad (26)$$

It is seen from (26) that

$$\mathbb{E}|\boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}w^o|^2 = \mathbb{E}|(\boldsymbol{u}_{k,i} - \boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}w^o|^2 + \sigma_{v,k}^2, \quad (27)$$

and, hence

$$\begin{aligned} J_{k,\min} &\triangleq \sigma_{v,k}^2 \\ &= \mathbb{E}|\boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}w^o|^2 - \mathbb{E}|(\boldsymbol{u}_{k,i} - \boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}w^o|^2 \\ &= \mathbb{E}|\boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}w^o|^2 - p\|w^o\|_{R_{u,k}}^2 - p\sigma_{\xi,k}^2\|w^o\|^2. \end{aligned} \quad (28)$$

After a sufficient number of iterations, and for sufficiently small step-sizes, the estimate $\boldsymbol{w}_{k,i}$ in (23) is expected to approach the global minimizer of (22), which we already know is w^o [37]. If we therefore replace w^o by $\boldsymbol{w}_{k,i-1}$ in (28) we get for $i \gg 1$:

$$J_{k,\min} \approx \mathbb{E}|\boldsymbol{e}_k(i)|^2 - p\|\boldsymbol{w}_{k,i-1}\|_{R_{u,k}}^2 - p\sigma_{\xi,k}^2\|\boldsymbol{w}_{k,i-1}\|^2. \quad (29)$$

It is still not possible to estimate $\sigma_{\xi,k}^2$ directly from (29) since the expression depends on $R_{u,k}$ and p. Suppose, as indicated earlier, that an estimate for p is available, say, $\widehat{p} < 1$. This is a reasonable assumption since in many situations in practice, it is generally known what percentage of the data is corrupted (as illustrated by the examples in Sec. V). From (20), we can then estimate $R_{u,k}$ by writing:

$$R_{u,k} \approx \frac{1}{1-\widehat{p}}R_{\bar{u},k} - \frac{\widehat{p}}{1-\widehat{p}}\sigma_{\xi,k}^2 I_M. \quad (30)$$

Substituting into (29) and solving for an estimate for $\sigma_{\xi,k}^2$ we obtain

$$\begin{aligned} \widehat{\sigma}_{\xi,k}^2 &\approx \frac{(1-\widehat{p})\mathbb{E}|\boldsymbol{e}_k(i)|^2 - \widehat{p}\|\boldsymbol{w}_{k,i-1}\|_{R_{\bar{u},k}}^2 - (1-\widehat{p})\sigma_{v,k}^2}{\widehat{p}(1-2\widehat{p})\|\boldsymbol{w}_{k,i-1}\|^2} \\ &\overset{(a)}{\approx} \frac{(1-\widehat{p})\mathbb{E}|\boldsymbol{e}_k(i)|^2 - \widehat{p}\|\boldsymbol{w}_{k,i-1}\|_{R_{\bar{u},k}}^2}{\widehat{p}(1-2\widehat{p})\|\boldsymbol{w}_{k,i-1}\|^2} \end{aligned} \quad (31)$$

where in (a) we assumed that the noise variance, $\sigma_{v,k}^2$, is sufficiently small compared to $\|\boldsymbol{w}_{k,i-1}\|_{R_{\bar{u},k}}^2$. Since $\mathbb{E}|\boldsymbol{e}_k(i)|^2$ and the diagonal matrix $R_{\bar{u},k}$ are unknown, we estimate them by means of smoothing filters from data realizations:

$$\widehat{\boldsymbol{R}}_{\bar{u},k}(i) = (1-\alpha_1)\widehat{\boldsymbol{R}}_{\bar{u},k}(i-1) + \alpha_1(\bar{\boldsymbol{u}}_{k,i}^*\bar{\boldsymbol{u}}_{k,i}) \odot I_M \quad (32)$$

$$\widehat{\boldsymbol{\sigma}}_{e,k}(i) = (1-\alpha_2)\widehat{\boldsymbol{\sigma}}_{e,k}(i-1) + \alpha_2|\boldsymbol{e}_k(i)|^2 \quad (33)$$

$$\boldsymbol{g}_k(i) = \frac{(1-\widehat{p})\widehat{\boldsymbol{\sigma}}_{e,k}(i) - \widehat{p}\|\boldsymbol{w}_{k,i-1}\|_{\widehat{\boldsymbol{R}}_{\bar{u},k}(i)}^2}{\widehat{p}(1-2\widehat{p})\|\boldsymbol{w}_{k,i-1}\|^2} \quad (34)$$

$$\widehat{\boldsymbol{\sigma}}_{\xi,k}^2(i) = (1-\alpha_3)\widehat{\boldsymbol{\sigma}}_{\xi,k}^2(i-1) + \alpha_3\boldsymbol{g}_k(i) \quad (35)$$

where $0 < \alpha_m \ll 1$, for $m = 1, 2, 3$. To prevent large fluctuations in estimating $\widehat{\boldsymbol{\sigma}}_{\xi,k}^2(i)$, we also use a smoothing filter for updating $\widehat{\boldsymbol{\sigma}}_{\xi,k}^2(i)$ in (35). Since the covariance matrix of the regressor is assumed to be diagonal, we use the Hadamard product in (32). It is noted that the algorithm does not require knowledge about the statistics of the data, e.g., the correlation matrices are assumed to be unknown. The resulting diffusion algorithm, henceforth called modified ATC (mATC), is listed in Algorithm 1.

It is clear from the listing of the algorithm that the operation of the diffusion strategy (37)–(38) is coupled to steps (39)–(42) for estimating $\sigma_{\xi,k}^2$. Proper operation of the algorithm requires a reliable estimate for $\sigma_{\xi,k}^2$. There are at least two ways to assist in this regard. One way is to use a small step-size α_3 in (42). A second way is to first run a few iterations of the standard diffusion algorithm without bias correction, i.e.,

$$\begin{aligned} \boldsymbol{\phi}_{k,i} &= \boldsymbol{w}_{k,i-1} + \mu_k\bar{\boldsymbol{u}}_{k,i}^*[\boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}\boldsymbol{w}_{k,i-1}] \\ \boldsymbol{w}_{k,i} &= \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \end{aligned} \quad (43)$$

Algorithm 1 Modified diffusion algorithm (mATC) for missing data

$$\boldsymbol{e}_k(i) \triangleq \boldsymbol{d}_k(i) - \bar{\boldsymbol{u}}_{k,i}\boldsymbol{w}_{k,i-1} \tag{36}$$

$$\boldsymbol{\phi}_{k,i} = (1+\mu_k p\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i-1))\boldsymbol{w}_{k,i-1} + \mu_k \bar{\boldsymbol{u}}^*_{k,i}\boldsymbol{e}_k(i) \tag{37}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{38}$$

$$\widehat{\boldsymbol{R}}_{\bar{u},k}(i) = (1-\alpha_1)\widehat{\boldsymbol{R}}_{\bar{u},k}(i-1) + \alpha_1(\bar{\boldsymbol{u}}^*_{k,i}\bar{\boldsymbol{u}}_{k,i}) \odot I_M \tag{39}$$

$$\widehat{\boldsymbol{\sigma}}_{e,k}(i) = (1-\alpha_2)\widehat{\boldsymbol{\sigma}}_{e,k}(i-1) + \alpha_2|\boldsymbol{e}_k(i)|^2 \tag{40}$$

$$\boldsymbol{g}_k(i) = \frac{(1-\widehat{p})\widehat{\boldsymbol{\sigma}}_{e,k}(i) - \widehat{p}\|\boldsymbol{w}_{k,i-1}\|^2_{\widehat{\boldsymbol{R}}_{\bar{u},k}(i)}}{\widehat{p}(1-2\widehat{p})\|\boldsymbol{w}_{k,i-1}\|^2} \tag{41}$$

$$\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i) = (1-\alpha_3)\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i-1) + \alpha_3\boldsymbol{g}_k(i) \tag{42}$$

and then switch to the censored version shown in the above listing. In the simulations section we illustrate both scenarios.

Remark 3: Since the estimate for $\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i)$ in (42) needs to be nonnegative, we can modify (41) to

$$\boldsymbol{g}_k(i) = \max\left\{\frac{(1-\widehat{p})\widehat{\boldsymbol{\sigma}}_{e,k}(i) - \widehat{p}\|\boldsymbol{w}_{k,i-1}\|^2_{\widehat{\boldsymbol{R}}_{\bar{u},k}(i)}}{\widehat{p}(1-2\widehat{p})\|\boldsymbol{w}_{k,i-1}\|^2}, 0\right\}. \tag{44}$$

IV. PERFORMANCE ANALYSIS

A. Error Dynamics

We associate with each agent the error vectors

$$\widetilde{\boldsymbol{\phi}}_{k,i} \triangleq w^o - \boldsymbol{\phi}_{k,i} \tag{45}$$

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i}. \tag{46}$$

Now, if we subtract w^o from both sides of (37) and (38) and replace $\boldsymbol{d}_k(i)$ by (26), we obtain

$$\begin{aligned}\widetilde{\boldsymbol{\phi}}_{k,i} =&(1+\mu_k p\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i-1))\widetilde{\boldsymbol{w}}_{k,i-1} - \mu_k p\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i-1)w^o - \\ &\mu_k\bar{\boldsymbol{u}}^*_{k,i}(\bar{\boldsymbol{u}}_{k,i}\widetilde{\boldsymbol{w}}_{k,i-1} + (\boldsymbol{u}_{k,i}-\boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}w^o + \boldsymbol{v}_k(i))\end{aligned} \tag{47}$$

$$\widetilde{\boldsymbol{w}}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\widetilde{\boldsymbol{\phi}}_{\ell,i}. \tag{48}$$

We collect the errors from across the network into the block vectors:

$$\widetilde{\boldsymbol{\phi}}_i \triangleq \mathrm{col}\{\widetilde{\boldsymbol{\phi}}_{1,i}, \cdots, \widetilde{\boldsymbol{\phi}}_{N,i}\} \tag{49}$$

$$\widetilde{\boldsymbol{w}}_i \triangleq \mathrm{col}\{\widetilde{\boldsymbol{w}}_{1,i}, \cdots, \widetilde{\boldsymbol{w}}_{N,i}\} \tag{50}$$

and note from (47)–(48) that they satisfy the following recursions:

$$\begin{aligned}\widetilde{\boldsymbol{\phi}}_i =&[I_{NM} - \mathcal{M}(\bar{\boldsymbol{\mathcal{R}}}_i - p\boldsymbol{\mathcal{K}}_{i-1}))]\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{M}\boldsymbol{s}_i - \\ &\mathcal{M}(\boldsymbol{\mathcal{R}}_{e,i} + p\boldsymbol{\mathcal{K}}_{i-1})w^o_e\end{aligned} \tag{51}$$

$$\widetilde{\boldsymbol{w}}_i = \mathcal{A}^T\widetilde{\boldsymbol{\phi}}_i \tag{52}$$

where we introduced the quantities:

$$w^o_e \triangleq \mathbb{1}_N \otimes w^o \tag{53}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{54}$$

$$\bar{\boldsymbol{\mathcal{R}}}_i \triangleq \mathrm{diag}\{\bar{\boldsymbol{u}}^*_{1,i}\bar{\boldsymbol{u}}_{1,i}, \bar{\boldsymbol{u}}^*_{2,i}\bar{\boldsymbol{u}}_{2,i}, \ldots, \bar{\boldsymbol{u}}^*_{N,i}\bar{\boldsymbol{u}}_{N,i}\} \tag{55}$$

$$\boldsymbol{\mathcal{R}}_{e,i} \triangleq \mathrm{diag}\left\{\{\bar{\boldsymbol{u}}^*_{k,i}(\boldsymbol{u}_{k,i}-\boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}\}_{k=1,\ldots,N}\right\} \tag{56}$$

$$\boldsymbol{\mathcal{K}}_{i-1} \triangleq \mathrm{diag}\left\{\widehat{\boldsymbol{\sigma}}^2_{\xi,1}(i-1)I_M, \ldots, \widehat{\boldsymbol{\sigma}}^2_{\xi,N}(i-1)I_M\right\} \tag{57}$$

$$\mathcal{M} \triangleq \mathrm{diag}\{\mu_1 I_M, \mu_2 I_M, \ldots, \mu_N I_N\} \tag{58}$$

$$\boldsymbol{s}_i \triangleq \mathrm{col}\{\bar{\boldsymbol{u}}^*_{1,i}\boldsymbol{v}_1(i), \ldots, \bar{\boldsymbol{u}}^*_{N,i}\boldsymbol{v}_N(i)\} \tag{59}$$

where the matrix A is left-stochastic, i.e., $A^T\mathbb{1}_M = \mathbb{1}_M$, with its (ℓ,k) entry equal to $a_{\ell k}$. If we now combine (51) and (52) we find that $\widetilde{\boldsymbol{w}}_i$ evolves according to the following dynamics:

$$\begin{aligned}\widetilde{\boldsymbol{w}}_i =& \mathcal{A}^T[I_{NM} - \mathcal{M}(\bar{\boldsymbol{\mathcal{R}}}_i - p\boldsymbol{\mathcal{K}}_{i-1})]\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{A}^T\mathcal{M}\boldsymbol{s}_i \\ &- \mathcal{A}^T\mathcal{M}(\boldsymbol{\mathcal{R}}_{e,i} + p\boldsymbol{\mathcal{K}}_{i-1})w^o_e\end{aligned} \tag{60}$$

From the definitions in (55), (56), and (59), we get

$$\mathbb{E}\,\boldsymbol{s}_i = 0 \tag{61}$$

$$\mathcal{S} \triangleq \mathbb{E}\,\boldsymbol{s}_i\boldsymbol{s}^*_i = \mathrm{diag}\{\sigma^2_{v,1}R_{\bar{u},1}, \ldots, \sigma^2_{v,N}R_{\bar{u},N}\} \tag{62}$$

$$\begin{aligned}\bar{\mathcal{R}} \triangleq& \mathbb{E}\bar{\boldsymbol{\mathcal{R}}}_i \\ =& \mathrm{diag}\{R_{\bar{u},1}, \ldots, R_{\bar{u},N}\} \\ =& (1-p)\mathcal{R} + p\,\mathrm{diag}\{\sigma^2_{\xi,1}I_M, \ldots, \sigma^2_{\xi,N}I_M\}\end{aligned} \tag{63}$$

$$\mathcal{R}_e \triangleq \mathbb{E}\boldsymbol{\mathcal{R}}_{e,i} = -p\,\mathrm{diag}\{\sigma^2_{\xi,1}I_M, \ldots, \sigma^2_{\xi,N}I_M\} \tag{64}$$

where

$$\mathcal{R} \triangleq \mathrm{diag}\{R_{u,1}, \ldots, R_{u,N}\} \tag{65}$$

and where we used the following result to compute $\mathcal{R}_e$ in (64):

$$\begin{aligned}&\mathbb{E}\bar{\boldsymbol{u}}^*_{k,i}(\boldsymbol{u}_{k,i}-\boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i} \\ &= \mathbb{E}\{[(I_M - \boldsymbol{F}_{k,i})\boldsymbol{u}^*_{k,i} + \boldsymbol{F}_{k,i}\boldsymbol{\xi}^*_{k,i}](\boldsymbol{u}_{k,i}-\boldsymbol{\xi}_{k,i})\boldsymbol{F}_{k,i}\} \\ &= \mathbb{E}\{\underbrace{(I_M - \boldsymbol{F}_{k,i})R_{u,k}\boldsymbol{F}_{k,i}}_{=0}\} - \sigma^2_{\xi,k}\mathbb{E}\,\boldsymbol{F}^2_{k,i} \\ &= -p\sigma^2_{\xi,k}I_M.\end{aligned} \tag{66}$$

B. Long-Term Approximations

It is clear from (39)–(42) that the operation of the diffusion strategy (37)–(38) is coupled with the estimation of the noise power $\sigma^2_{\xi,k}$. This is because the estimate $\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i)$ in (42) is dependent on $\boldsymbol{w}_{k,i-1}$. This coupling makes the performance analysis of the algorithm rather challenging. Since we are mainly interested in assessing the performance of the solution in the infinite-horizon after sufficient iterations have elapsed, and after the algorithm has been given sufficient time to learn, we are going to proceed from this point onwards under the assumption that $i \gg 1$ and that the smoothing filters (39)–(42) have approached steady-state operation. Specifically, in steady state, i.e., for $i \to \infty$, taking the expectation of both

sides of Eq. (39) and considering $(1-\alpha_1)^i \to 0$ for $i \gg 1$, we obtain:

$$\begin{aligned}\mathbb{E}\widehat{\boldsymbol{R}}_{\bar{u},k}(i) &= \sum_{j=0}^{i} \alpha_1(1-\alpha_1)^{i-j}\mathbb{E}(\bar{\boldsymbol{u}}_{k,j}^{*}\bar{\boldsymbol{u}}_{k,j}) \odot I_M \\ &= \alpha_1 \frac{1-(1-\alpha_1)^{i+1}}{1-(1-\alpha_1)} R_{\bar{u},k} \odot I_M \\ &\to R_{\bar{u},k}, \quad i \gg 1 \end{aligned} \tag{67}$$

so that $\widehat{\boldsymbol{R}}_{\bar{u},k}$ tends on average to the true value $R_{\bar{u},k}$. Similarly, for Eqs. (40)–(42), they approach steady-state operation with

$$\mathbb{E}\widehat{\boldsymbol{\sigma}}_{e,k}^2(i) \to \mathbb{E}|\boldsymbol{e}_k(i)|^2, \quad i \gg 1 \tag{68}$$

$$\mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,k}^2(i) \to \mathbb{E}\boldsymbol{g}_k(i), \quad i \gg 1. \tag{69}$$

We now estimate $\mathbb{E}\boldsymbol{g}_k(i)$ by employing a first-order Taylor series approximation. Let $\boldsymbol{z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]$ be a real random vector with mean $\mathbb{E}\boldsymbol{z} = [\mathbb{E}\boldsymbol{z}_1, \ldots, \mathbb{E}\boldsymbol{z}_n]$. The first-order Taylor series expansion of a differentiable real function $f(\boldsymbol{z})$ about the mean $\mathbb{E}\boldsymbol{z}$ can be expressed as follows [58, p. 241]– [59, p. 295]:

$$f(\boldsymbol{z}) \approx f(\mathbb{E}\boldsymbol{z}) + \sum_{k=1}^{n} \frac{\partial}{\partial \boldsymbol{z}_k} f(\boldsymbol{z}_k)|_{\boldsymbol{z}_k = \mathbb{E}\boldsymbol{z}_k}(\boldsymbol{z}_k - \mathbb{E}\boldsymbol{z}_k). \tag{70}$$

If we evaluate the expectation of both sides of (70) with respect to the random vector $\boldsymbol{z}$, we get the follwoing approximation:

$$\mathbb{E}f(\boldsymbol{z}) \approx f(\mathbb{E}\boldsymbol{z}). \tag{71}$$

Now consider a function of the ratio of two random variables as $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}/\boldsymbol{y}$, and assume that $\boldsymbol{y}$ has nonzero mean. From (71), we can write

$$\mathbb{E}f(\boldsymbol{x}, \boldsymbol{y}) \approx \frac{\mathbb{E}\boldsymbol{x}}{\mathbb{E}\boldsymbol{y}}. \tag{72}$$

To approximate $\mathbb{E}\boldsymbol{g}_k(i)$, we apply (72) to (41) as follows:

$$\mathbb{E}\boldsymbol{g}_k(i) \approx \frac{(1-\widehat{p})\mathbb{E}\widehat{\boldsymbol{\sigma}}_{k,e}^2(i) - \widehat{p}\,\mathbb{E}\left(\|\boldsymbol{w}_{k,i}\|^2_{\widehat{\boldsymbol{R}}_{\bar{u},k}(i)}\right)}{\widehat{p}(1-2\widehat{p})\mathbb{E}\|\boldsymbol{w}_{k,i}\|^2}. \tag{73}$$

For tractability, we assume that, in steady-state:

$$\mathbb{E}\left(\|\boldsymbol{w}_{k,i}\|^2_{\widehat{\boldsymbol{R}}_{\bar{u},k}(i)}\right) \approx \|w^o\|^2_{R_{\bar{u},k}}. \tag{74}$$

so that

$$\begin{aligned}\mathbb{E}\boldsymbol{g}_k(i) &\approx \frac{(1-\widehat{p})\mathbb{E}|\boldsymbol{e}_k(i)|^2 - \widehat{p}\|w^o\|^2_{R_{\bar{u},k}}}{\widehat{p}(1-2\widehat{p})\|w^o\|^2} \\ &\overset{(a)}{\approx} \frac{(1-\widehat{p})(\mathbb{E}|\boldsymbol{e}_k(i)|^2 - \widehat{p}\|w^o\|^2_{R_{u,k}}) - \widehat{p}^2\sigma_{\xi,k}^2\|w^o\|^2}{\widehat{p}(1-2\widehat{p})\|w^o\|^2} \\ &\overset{(b)}{\approx} \frac{(1-\widehat{p})(\sigma_{v,k}^2 + \widehat{p}\sigma_{\xi,k}^2\|w^o\|^2) - \hat{p}^2\sigma_{\xi,k}^2\|w^o\|^2}{\widehat{p}(1-2\widehat{p})\|w^o\|^2} \\ &\overset{(c)}{\approx} \sigma_{\xi,k}^2 \end{aligned} \tag{75}$$

where in (a) we replace $R_{\bar{u},k}$ from (20), in (b) we use the relation from (28), and in (c) we assume that the term $(1-\widehat{p})\sigma_{v,k}^2$ is sufficiently small compared to $\widehat{p}(1-2\widehat{p})\sigma_{\xi,k}^2\|w^o\|^2$. Therefore, in steady state, we set

$$\mathcal{K} \triangleq \lim_{i\to\infty} \mathbb{E}\boldsymbol{\mathcal{K}}_i \approx \operatorname{diag}\left\{\sigma_{\xi,1}^2 I_M, \ldots, \sigma_{\xi,N}^2 I_M\right\}. \tag{76}$$

C. Mean Stability Analysis

First note that, in steady state, from (64) and (76) we have

$$\lim_{i\to\infty} \mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i} + p\boldsymbol{\mathcal{K}}_{i-1}) \approx 0. \tag{77}$$

Now, since the variables $\boldsymbol{u}_{k,i}$ and $\boldsymbol{\xi}_{k,i}$ are temporally white and spatially independent, then the error vectors $\widetilde{\boldsymbol{w}}_{\ell,j}$ are independent of $\boldsymbol{u}_{k,i}$ and $\boldsymbol{\xi}_{k,i}$ for all j if $k \neq \ell$ and for $k = \ell$ if $j \leq i-1$. Therefore, taking expectation of both sides of (60) for large enough i gives under the long-term approximations of the previous section:

$$\mathbb{E}\widetilde{\boldsymbol{w}}_i = \mathcal{A}^T[I_{NM} - (1-p)\mathcal{M}\mathcal{R}]\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}, \quad i \gg 1. \tag{78}$$

This recursion is stable if the step sizes are sufficiently small and satisfy

$$0 < \mu_k < \frac{2}{(1-p)\lambda_{\max}(R_{u,k})} \tag{79}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of its matrix argument. In this case, the estimator becomes asymptotically unbiased, i.e., $\lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_i = 0$.

D. Mean Variance Analysis

We rewrite (60) more compactly as

$$\widetilde{\boldsymbol{w}}_i = \boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{G}\boldsymbol{s}_i - \boldsymbol{\mathcal{D}}_i w_e^o \tag{80}$$

where

$$\boldsymbol{\mathcal{B}}_i \triangleq \mathcal{A}^T[I_{NM} - \mathcal{M}(\bar{\boldsymbol{\mathcal{R}}}_i - p\boldsymbol{\mathcal{K}}_{i-1})] \tag{81}$$

$$\boldsymbol{\mathcal{D}}_i \triangleq \mathcal{A}^T\mathcal{M}[\boldsymbol{\mathcal{R}}_{e,i} + p\boldsymbol{\mathcal{K}}_{i-1}] \tag{82}$$

$$\mathcal{G} \triangleq \mathcal{A}^T\mathcal{M}. \tag{83}$$

The mean-square error analysis of the algorithm relies on evaluating a weighted variance of the error vector. Let Σ denote an arbitrary nonnegative-definite matrix that we are free to choose. We express the weighted square measures on both sides of (80) as

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_i\|^2_\Sigma =& \|\boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1} - \mathcal{G}\boldsymbol{s}_i - \boldsymbol{\mathcal{D}}_i w_e^o\|^2_\Sigma \\ =& \widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i^*\Sigma\boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1} + \boldsymbol{s}_i^*\mathcal{G}^T\Sigma\mathcal{G}\boldsymbol{s}_i + w_e^{o*}\boldsymbol{\mathcal{D}}_i^*\Sigma\boldsymbol{\mathcal{D}}_i w_e^o \\ &- \widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i^*\Sigma\mathcal{G}\boldsymbol{s}_i - \boldsymbol{s}_i^*\mathcal{G}^T\Sigma\boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1} \\ &- \widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i^*\Sigma\boldsymbol{\mathcal{D}}_i w_e^o - w_e^{o*}\boldsymbol{\mathcal{D}}_i^*\Sigma\boldsymbol{\mathcal{B}}_i\widetilde{\boldsymbol{w}}_{i-1} \\ &+ \boldsymbol{s}_i^*\mathcal{G}\Sigma\boldsymbol{\mathcal{D}}_i w_e^o + w_e^{o*}\boldsymbol{\mathcal{D}}_i^*\Sigma\mathcal{G}\boldsymbol{s}_i. \end{aligned} \tag{84}$$

We now compute the expectation of both sides of (84) in steady state. Since $\lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_i = 0$, $\mathbb{E}\boldsymbol{s}_i = 0$, and $\widetilde{\boldsymbol{w}}_{i-1}$ and $\boldsymbol{s}_i$ are independent of each other, we get the following results:

$$\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i^*\Sigma\mathcal{G}\boldsymbol{s}_i = 0 \tag{85}$$

$$\mathbb{E}w_e^{o*}\boldsymbol{\mathcal{D}}_i^*\Sigma\mathcal{G}\boldsymbol{s}_i = 0 \tag{86}$$

$$\begin{aligned}\lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i\Sigma\boldsymbol{\mathcal{D}}_i w_e^o &= \lim_{i\to\infty}\mathbb{E}[\mathbb{E}(\widetilde{\boldsymbol{w}}_{i-1}^*\boldsymbol{\mathcal{B}}_i^*\Sigma\boldsymbol{\mathcal{D}}_i|_{\widetilde{\boldsymbol{w}}_{i-1}})]w_e^o \\ &= \lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}^*[\mathbb{E}(\boldsymbol{\mathcal{B}}_i^*\Sigma\boldsymbol{\mathcal{D}}_i|_{\widetilde{\boldsymbol{w}}_{i-1}})]w_e^o \\ &= \lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}^*[\mathbb{E}\mathcal{A}\Sigma\boldsymbol{\mathcal{D}}_i]w_e^o + O(\mathcal{M}^2) \\ &\approx \lim_{i\to\infty}\mathbb{E}\widetilde{\boldsymbol{w}}_{i-1}^*[\mathbb{E}\mathcal{A}\Sigma\boldsymbol{\mathcal{D}}_i]w_e^o \\ &\approx 0, \end{aligned} \tag{87}$$

since from (77)

$$\mathbb{E}\boldsymbol{\mathcal{D}}_i \approx 0, \quad i \gg 1. \tag{88}$$

Noting that the cross-terms are either zero or negligible under expectation, we conclude from (80)–(87) that

$$\begin{aligned}\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\Sigma}^2 = \lim_{i\to\infty} \big[&\mathbb{E}(\widetilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{\mathcal{B}}_i^* \Sigma \boldsymbol{\mathcal{B}}_i \widetilde{\boldsymbol{w}}_{i-1}) + \mathbb{E}(\boldsymbol{s}_i^* \mathcal{G}^T \Sigma \mathcal{G} \boldsymbol{s}_i) \\ &+ \mathbb{E}(w_e^{o*} \boldsymbol{\mathcal{D}}_i^* \Sigma \boldsymbol{\mathcal{D}}_i w_e^o)\big].\end{aligned} \tag{89}$$

We now evaluate the terms that appear on the right-hand side of (89). The first term can be written as

$$\lim_{i\to\infty} \mathbb{E}(\widetilde{\boldsymbol{w}}_{i-1}^* \boldsymbol{B}_i^* \Sigma \boldsymbol{B}_i \widetilde{\boldsymbol{w}}_{i-1}) = \lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 \tag{90}$$

where

$$\begin{aligned}\Sigma' &\triangleq \lim_{i\to\infty} \mathbb{E}(\boldsymbol{\mathcal{B}}_i^* \Sigma \boldsymbol{\mathcal{B}}_i) \\ &= \mathcal{A}\Sigma\mathcal{A}^T - (1-p)\mathcal{R}\mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T - (1-p)\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\mathcal{R} \\ &\quad + O(\mathcal{M}^2),\end{aligned} \tag{91}$$

in view of the fact that

$$\lim_{i\to\infty} \mathbb{E}(\bar{\boldsymbol{\mathcal{R}}}_i - p\boldsymbol{\mathcal{K}}_{i-1}) = (1-p)\mathcal{R} \tag{92}$$

with $\mathcal{R}$ defined by (65). For sufficiently small step-sizes, the term $O(\mathcal{M}^2)$ in (91) is negligible and we set

$$\Sigma' = \mathcal{A}\Sigma\mathcal{A}^T - (1-p)\mathcal{R}\mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T - (1-p)\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\mathcal{R}. \tag{93}$$

The second term on the right-hand side of (89) can be calculated as

$$\mathbb{E}(\boldsymbol{s}_i^* \mathcal{G}^T \Sigma \mathcal{G} \boldsymbol{s}_i) = \mathbb{E}\,\mathrm{Tr}(\mathcal{G}^T \Sigma \mathcal{G} \boldsymbol{s}_i \boldsymbol{s}_i^*) = \mathrm{Tr}(\mathcal{G}^T \Sigma \mathcal{G} \mathcal{S}). \tag{94}$$

Now it is clear from expression (56) that $\boldsymbol{\mathcal{R}}_{e,i}$ depends on random variables that are available at time i, while expression (57) shows that $\boldsymbol{\mathcal{K}}_{i-1}$ depends on different random variables up to time $i-1$. Therefore, $\boldsymbol{\mathcal{R}}_{e,i}$ and $\boldsymbol{\mathcal{K}}_{i-1}$ are independent of each other. Then, the last term on the right-hand side of (89) can be computed as follows:

$$\begin{aligned}&\lim_{i\to\infty} \mathbb{E}(w_e^{o*} \boldsymbol{\mathcal{D}}_i^* \Sigma \boldsymbol{\mathcal{D}}_i w_e^o) \\ &\quad = \lim_{i\to\infty} \big[2pw_e^{o*}\mathrm{Re}\{\mathbb{E}\boldsymbol{\mathcal{R}}_{e,i}^* \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{\mathcal{K}}_{i-1}^*\}w_e^o \\ &\qquad + w_e^{o*}\mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i}^* \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{\mathcal{R}}_{e,i})w_e^o \\ &\qquad + p^2 w_e^{o*}\mathbb{E}(\boldsymbol{\mathcal{K}}_{i-1}^* \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{\mathcal{K}}_{i-1})w_e^o\big] \\ &\quad \overset{(a)}{=} -2w_e^{o*}\mathcal{R}_e\mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\mathcal{R}_e w_e^o \\ &\qquad + \lim_{i\to\infty} w_e^{o*}\mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i}^* \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{\mathcal{R}}_{e,i})w_e^o \\ &\qquad + \lim_{i\to\infty} p^2 w_e^{o*}\mathbb{E}[\boldsymbol{\mathcal{K}}_{i-1}^* \mathcal{M}\mathcal{A}\Sigma\mathcal{A}^T\mathcal{M}\boldsymbol{\mathcal{K}}_{i-1}]w_e^o.\end{aligned} \tag{95}$$

In (a), we used the independence of $\boldsymbol{\mathcal{R}}_{e,i}$ and $\boldsymbol{\mathcal{K}}_{i-1}$ and the fact from (77) that $\mathbb{E}\boldsymbol{\mathcal{R}}_{e,i} = -p\mathbb{E}\boldsymbol{\mathcal{K}}_{i-1},\ i \gg 1$. In order to obtain a more compact representation for the variance relation, we use the following vector notation:

$$\sigma = \mathrm{vec}(\Sigma), \quad \sigma' = \mathrm{vec}(\Sigma'), \tag{96}$$

where the vec operator vectorizes a matrix by placing its columns on top of each other. We also use the following Kroneceker product properties [37]:

$$\begin{aligned}\mathrm{vec}(ABC) &= (C^T \otimes A)\mathrm{vec}(B) \\ \mathrm{Tr}(AB) &= [\mathrm{vec}(B^T)]^T \mathrm{vec}(A).\end{aligned} \tag{97}$$

Then, from (93) we can write

$$\sigma' = \mathcal{F}\sigma, \tag{98}$$

where

$$\mathcal{F} \triangleq \mathcal{A}\otimes\mathcal{A} - (1-p)\mathcal{A}\otimes\mathcal{R}\mathcal{M}\mathcal{A} - (1-p)\mathcal{R}^T\mathcal{M}\mathcal{A}\otimes\mathcal{A}. \tag{99}$$

Likewise, we have

$$\mathrm{Tr}(\mathcal{G}^T\Sigma\mathcal{G}\mathcal{S}) = \mathrm{Tr}(\Sigma\mathcal{G}\mathcal{S}\mathcal{G}^T) = [\mathrm{vec}(\mathcal{G}\mathcal{S}^T\mathcal{G}^T)]^T\sigma. \tag{100}$$

In a similar manner, we can express the right-hand side of (95) using the vec operator as follows:

$$\begin{aligned}&\lim_{i\to\infty} \mathbb{E}(w_e^{o*} \boldsymbol{\mathcal{D}}_i^* \Sigma \boldsymbol{\mathcal{D}}_i w_e^o) \\ &\quad = -2\mathrm{Tr}(\Sigma\mathcal{A}^T\mathcal{M}\mathcal{R}_e w_e^o w_e^{o*}\mathcal{R}_e\mathcal{M}\mathcal{A}) \\ &\qquad + \mathrm{Tr}(\Sigma\mathcal{A}^T\mathcal{M}\lim_{i\to\infty}\mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i} w_e^o w_e^{o*}\boldsymbol{\mathcal{R}}_{e,i}^*)\mathcal{M}\mathcal{A}) \\ &\qquad + p^2\mathrm{Tr}(\Sigma\mathcal{A}^T\mathcal{M}\lim_{i\to\infty}\mathbb{E}(\boldsymbol{\mathcal{K}}_{i-1} w_e^o w_e^{o*}\boldsymbol{\mathcal{K}}_{i-1}^T)\mathcal{M}\mathcal{A}) \\ &\quad = -2[\mathrm{vec}(\mathcal{A}^T\mathcal{M}\mathcal{R}_e w_e^o w_e^{o*}\mathcal{R}_e\mathcal{M}\mathcal{A})^T]^T\sigma \\ &\qquad + [\mathrm{vec}(\mathcal{A}^T\mathcal{M}\lim_{i\to\infty}\mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i} w_e^o w_e^{o*}\boldsymbol{\mathcal{R}}_{e,i}^*)\mathcal{M}\mathcal{A})^T]^T\sigma \\ &\qquad + p^2[\mathrm{vec}(\mathcal{A}^T\mathcal{M}\lim_{i\to\infty}\underbrace{\mathbb{E}(\boldsymbol{\mathcal{K}}_{i-1} w_e^o w_e^{o*}\boldsymbol{\mathcal{K}}_{i-1}^T)}_{\triangleq\Pi}\mathcal{M}\mathcal{A})^T]^T\sigma.\end{aligned} \tag{101}$$

It is noted that the quantity $\mathbb{E}(\boldsymbol{\mathcal{R}}_{e,i} w_e^o w_e^{o*}\boldsymbol{\mathcal{R}}_{e,i}^*)$, in general, does not have a closed form expression. We can approximate it by means of ensemble averaging.

The last term on the right-hand side of (101) can be approximated as follows. Let $\Omega \triangleq w_e^o w_e^{o*}$ be a Hermitian block matrix, whose (n,m) block is given by

$$\Omega_{n,m} = w^o w^{o*}. \tag{102}$$

The (n,m)-th block of Π can be obtained as follows:

$$\begin{aligned}\Pi_{n,m} &= \mathbb{E}(\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\Omega_{n,m}\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i-1)) \\ &= \Omega_{n,m}\mathbb{E}(\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i-1)).\end{aligned} \tag{103}$$

From (42), we can write

$$\begin{aligned}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i) =&(1-\alpha_3)^2\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i-1) \\ &+ \alpha_3^2\boldsymbol{g}_n(i)\boldsymbol{g}_m(i) \\ &+ \alpha_3(1-\alpha_3)[\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\boldsymbol{g}_m(i) \\ &+ \widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i-1)\boldsymbol{g}_n(i)].\end{aligned} \tag{104}$$

In steady state, we have $\mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i) = \mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i-1)$. We further benefit from the following approximations for $i \gg 1$:

$$\begin{aligned}\mathbb{E}|\boldsymbol{e}_n(j)|^2|\boldsymbol{e}_m(i)|^2 &\approx \mathbb{E}|\boldsymbol{e}_n(j)|^2\mathbb{E}|\boldsymbol{e}_m(i)|^2 \\ \mathbb{E}\|\boldsymbol{w}_{n,i-1}\|^2\|\boldsymbol{w}_{m,i-1}\|^2 &\approx \|w^o\|^4 \\ \mathbb{E}\|\boldsymbol{w}_{n,i-1}\|_{\widehat{\boldsymbol{R}}_{\bar{u},n}(i)}^2\|\boldsymbol{w}_{m,i-1}\|_{\widehat{\boldsymbol{R}}_{\bar{u},m}(i)}^2 &\approx \|w^o\|_{R_{\bar{u},n}}^2\|w^o\|_{R_{\bar{u},m}}^2.\end{aligned} \tag{105}$$

We again use a first-order Taylor series expansion to approximate the following terms as:

$$\mathbb{E}\boldsymbol{g}_n(i)\boldsymbol{g}_m(i) \approx \mathbb{E}\boldsymbol{g}_n(i)\mathbb{E}\boldsymbol{g}_m(i) \approx \sigma_{\xi,n}^2\sigma_{\xi,m}^2 \tag{106}$$

$$\mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\boldsymbol{g}_m(i) \approx \mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i-1)\mathbb{E}\boldsymbol{g}_m(i) \approx \sigma_{\xi,n}^2\sigma_{\xi,m}^2. \tag{107}$$

Using the above expressions and (104), we obtain

$$(2\alpha_3-\alpha_3^2)\mathbb{E}\widehat{\boldsymbol{\sigma}}_{\xi,n}^2(i)\widehat{\boldsymbol{\sigma}}_{\xi,m}^2(i) \approx (2\alpha_3-\alpha_3^2)\sigma_{\xi,n}^2\sigma_{\xi,m}^2. \tag{108}$$

Hence, in steady state, the matrix Π is approximated as

$$\Pi \approx \mathcal{K}w^o w^{o*}\mathcal{K} \tag{109}$$

and in steady state, expression (95) is approximated as

$$\lim_{i\to\infty}\mathbb{E}(w_e^{o*}\boldsymbol{\mathcal{D}}_i^*\Sigma\boldsymbol{\mathcal{D}}_i w_e^o) \approx [\mathrm{vec}(\mathcal{Z}^T)]^T\sigma \tag{110}$$

where

$$\begin{aligned}\mathcal{Z} \triangleq &-\mathcal{A}^T\mathcal{M}\mathcal{R}_e w_e^o w_e^{o*}\mathcal{R}_e\mathcal{M}\mathcal{A}\\ &+\mathcal{A}^T\mathcal{M}\mathbb{E}(\boldsymbol{R}_{e,i}w_e^o w_e^{o*}\boldsymbol{R}_{e,i}^*)\mathcal{M}\mathcal{A}.\end{aligned} \tag{111}$$

Referring back to (89), and using the notation $\|x\|_\sigma^2$ interchangeably with $\|x\|_\Sigma^2$, we get

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_\sigma = \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma} + [\mathrm{vec}(\mathcal{Z}^T+\mathcal{Y}^T)]^T\sigma \tag{112}$$

where

$$\mathcal{Y} \triangleq \mathcal{G}\mathcal{S}\mathcal{G}^T. \tag{113}$$

It follows that

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{(I-\mathcal{F})\sigma} = [\mathrm{vec}(\mathcal{Z}^T+\mathcal{Y}^T)]^T\sigma. \tag{114}$$

Based on the variance relation (114), different quantities can be computed. For example, we can evaluate the network and individual mean-square deviation (MSDs), respectively, defined as

$$\begin{aligned}\mathrm{MSD}^{\mathrm{network}} &\triangleq \lim_{i\to\infty}\frac{1}{N}\sum_{k=1}^N\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\frac{1}{N}}^2\\ \mathrm{MSD}_k &\triangleq \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = \lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\mathcal{I}_k}^2\end{aligned} \tag{115}$$

where $\mathcal{I}_k \triangleq \mathrm{diag}\{0,\ldots,0,I_M,0,\ldots,0\}$, with the identity matrix appearing in the $k-$th block location.

In order to derive the network MSD from variance relation (114), we select the weighting vector σ such that

$$(I-\mathcal{F})\sigma = \frac{1}{N}\mathrm{vec}(I_{NM}). \tag{116}$$

Then, the network MSD can be calculated as

$$\mathrm{MSD}^{\mathrm{network}} = \frac{1}{N}[\mathrm{vec}(\mathcal{Z}^T+\mathcal{Y}^T)]^T(I-\mathcal{F})^{-1}\mathrm{vec}(I_{NM}). \tag{117}$$

Likewise, the individual MSD can be computed as

$$\mathrm{MSD}_k = [\mathrm{vec}(\mathcal{Z}^T+\mathcal{Y}^T)]^T(I-\mathcal{F})^{-1}\mathrm{vec}(\mathcal{I}_k). \tag{118}$$

V. Design Examples

In this section, we evaluate the performance of the proposed diffusion algorithm through computer simulations. In all simulations, we consider the connected network of 7 agents shown in Fig. 1 and employ the uniform combination rule $a_{\ell,k} = 1/|\mathcal{N}_k|$ [57] and set the step-sizes across the agents to a uniform value, $\mu_k = \mu$. To evaluate the MSD, we run 400 experiments and compute the average MSD across these experiments for different approaches.

Some distributed techniques that rely on EM techniques, of the form studied in [60]–[62], have been proposed for useful but different applications over sensor networks, such as clustering or density estimation, but not for missing data scenarios considered in this study. We therefore compare our solution against centralized processing techniques. We also mentioned earlier that the missing data model considered in this paper is more general than the models investigated in the missing data literature. Therefore, to assess the performance of the proposed approach against existing techniques, first we try to detect the missing position and then employ existing techniques that require knowledge of these positions. Note that the suggested approach in this section may not be optimal. It may be possible to improve the performance of exiting approaches if we pursue *joint* detection of the missing position and the estimation of the unknown parameter vector rather than follow the two-step procedure used in this section.

Suppose the $j-$th component of the regressor $\boldsymbol{u}_{k,i}$ at the $k-$th node at time i, denoted by $\boldsymbol{u}_{k,i}^j$, is missing. For such a scenario, $\boldsymbol{F}_{k,i} = \mathrm{diag}\{0,\ldots,\boldsymbol{f}_{k,i}^j,\ldots,0\}$. We define two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ as follows:

$$\mathcal{H}_0 : \bar{\boldsymbol{u}}_{k,i}^j = \boldsymbol{u}_{k,i}^j \tag{119}$$

$$\mathcal{H}_1 : \bar{\boldsymbol{u}}_{k,i}^j = \boldsymbol{\xi}_{k,i}^j. \tag{120}$$

Since there are prior beliefs about $\mathcal{H}_0$ and $\mathcal{H}_1$, we consider the Bayesian hypothesis testing approach to detect the missing position [63]. That is, we should decide $\mathcal{H}_1$ ($\boldsymbol{f}_{k,i}^j = 1$) if the likelihood ratio (LLR) test is larger than a threshold γ as [64, Ch. 3]:

$$\mathrm{LLR} \triangleq \frac{p_{\mathcal{H}_1}(\bar{\boldsymbol{u}}_{k,i}^j)}{p_{\mathcal{H}_0}(\bar{\boldsymbol{u}}_{k,i}^j)} \overset{\mathcal{H}_1}{>} \gamma \triangleq \frac{(1-\widehat{p})}{\widehat{p}} \tag{121}$$

where $p_{\mathcal{H}_0}(\bar{\boldsymbol{u}}_{k,i}^j)$ and $p_{\mathcal{H}_1}(\bar{\boldsymbol{u}}_{k,i}^j)$ are the probability density functions under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

In general, evaluating the decision rule in (121) might be difficult. In this example, we assume that $\boldsymbol{u}_{k,i}^j$ and $\boldsymbol{\xi}_{k,i}^j$ are normally distributed. Then,

$$p_{\mathcal{H}_0}(\bar{\boldsymbol{u}}_{k,i}^j) = \frac{1}{\sqrt{2\pi R_{u,k}(j,j)}}\exp\left[-\frac{|\bar{\boldsymbol{u}}_{k,i}^j|^2}{2R_{u,k}(j,j)}\right] \tag{122}$$

$$p_{\mathcal{H}_1}(\bar{\boldsymbol{u}}_{k,i}^j) = \frac{1}{\sqrt{2\pi\sigma_{\xi,k}^2}}\exp\left[-\frac{|\bar{\boldsymbol{u}}_{k,i}^j|^2}{2\sigma_{\xi,k}^2}\right]. \tag{123}$$

The LLR is accordingly given by

$$\mathrm{LLR} = \sqrt{\frac{R_{u,k}(j,j)}{\sigma_{\xi,k}^2}}\exp\left[\frac{|\bar{\boldsymbol{u}}_{k,i}^j|^2}{2R_{u,k}(j,j)} - \frac{|\bar{\boldsymbol{u}}_{k,i}^j|^2}{2\sigma_{\xi,k}^2}\right] \tag{124}$$

and the decision rule can be expressed as follows:

$$|\bar{\boldsymbol{u}}_{k,i}^{j}|^2 \left[\frac{1}{R_{u,k}(j,j)} - \frac{1}{\sigma_{\xi,k}^2}\right] \overset{\mathcal{H}_1}{>} \log\left(\frac{(1-\widehat{p})^2\sigma_{\xi,k}^2}{\widehat{p}^2 R_{u,k}(j,j)}\right). \quad (125)$$

It is still seen that for evaluating the decision rule (125), we need to know the variances $R_{u,k}(j,j)$ and $\sigma_{\xi,k}^2$. To make the detection approach feasible, we assume that a good approximation for the ratio $R_{u,k}(j,j)/\sigma_{\xi,k}^2 = r_k$ is available. Note that for the proposed mATC algorithm, we do not need to know the ratio r_k. From (20), we have

$$(1-\widehat{p})R_{u,k}(j,j) + \widehat{p}\sigma_{\xi,k}^2 = R_{\bar{u},k}(j,j) \approx \frac{1}{M_k}\sum_{i=1}^{M_k} \bar{\boldsymbol{u}}_{k,i}^{j}\bar{\boldsymbol{u}}_{k,i}^{*j} \quad (126)$$

where M_k is the number of measurements collected by node k. Substituting $R_{u,k}(j,j)$ by $r_k\sigma_{\xi,k}^2$ in (126), we can approximate the variance $\sigma_{\xi,k}^2$, denoted by $\check{\sigma}_{\xi,k}^2$, as

$$\check{\sigma}_{\xi,k}^2 = \frac{1}{M_k((1-\widehat{p})r_k + \widehat{p})}\sum_{i=1}^{M_k} \bar{\boldsymbol{u}}_{k,i}^{j}\bar{\boldsymbol{u}}_{k,i}^{*j}. \quad (127)$$

The variance $R_{u,k}(j,j)$ can be estimated as

$$\widehat{R}_{u,k}(j,j) = r_k\check{\sigma}_{\xi,k}^2. \quad (128)$$

We note that the estimate of the variance $\sigma_{\xi,k}^2$ in (127) needs the ratio r_k to be known and also the estimate is based on a batch processing.

Once the missing positions have been identified, we then apply two state-of-the-art centralized techniques to centrally estimate the unknown vector w. In the first approach, the detected missing position is filled by the mean of the data and then a least-squares construction is applied (Imput-LS). For details on this approach, the reader may refer to [65]. In the second approach, we obtain the maximum likelihood estimator (MLE) assuming known distributions for the regressor and perturbation $\boldsymbol{v}_k(i)$. The details of the MLE for missing data can be found in [4], [21], [65], [66].

Remark 4: The two-step approach above may not be an optimal implementation for the centralized solution. It may be possible to develop more efficient centralized algorithms based, for example, on the EM algorithm and mixture models [24]–[26], [67], [68]. It is noted that if a mixture model for the missing data in (3) is considered, then a large number of components may be needed.

In the first simulation, we assume a Gaussian distribution for the process noise, $\boldsymbol{v}_k(i) \sim \mathcal{N}(0, \sigma_{v,k}^2)$ with $\sigma_{v,k}^2 = 0.01$. The regressor $\boldsymbol{u}_{k,i}$ has Gaussian distribution with diagonal covariance matrix, $R_{u,k} = \text{diag}\{1, 1.6, 0.8, 0.95, 1.2\}$. The probability of missing is set to $p = 0.3$ or $p = 0.4$, which is assumed to be known in advance. We set the unknown vector $w^o = [1,\ -0.5,\ 1.2,\ 0.4,\ 1.5]^T$. We also assume that $\boldsymbol{\xi}_{k,i} \sim \mathcal{N}(0, \sigma_{\xi,k}^2)$ with variances $\sigma_{\xi,1}^2 = 0.02$, $\sigma_{\xi,2}^2 = 0.44$, $\sigma_{\xi,3}^2 = 0.04$, $\sigma_{\xi,4}^2 = 0.09$, $\sigma_{\xi,5}^2 = 0.15$, $\sigma_{\xi,6}^2 = 0.26$, and $\sigma_{\xi,7}^2 = 0.13$. The step size μ is set to 0.04 for every node. In the simulation, we assume that the first component of the regressor is missing. In this simulation, the recursions for estimating the variance $\sigma_{\xi,k}^2$, i.e., Eqs. (39)–(42), start after

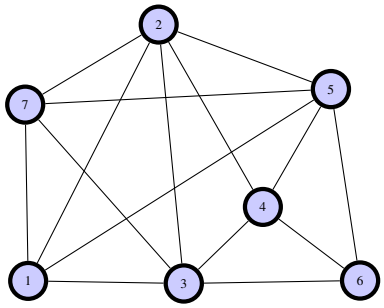


Fig. 1. Topology of the network used in the simulations.

50 iterations ($i = 50$) of Eqs. (37)–(38). We set $\alpha_1 = \alpha_2 = \alpha_3 = 0.01$.

Figures 2(a) and 2(b) show the MSD learning curves for different approaches for $p = 0.3$ and $p = 0.4$. As it is observed, the mATC shows a promising performance compared to centralized approaches. As the probability of missing increases, the proposed mATC considerably outperforms other approaches. The reason is that the error of missing position detection will increase as the probability of missing increases. Therefore, the performance of centralized approaches will be degraded more with increasing the probability of missing p.

In the next sections, we consider two specific applications.

A. Household Consumption

In this section, we evaluate the performance of the proposed algorithm for a household consumption application. Household consumption depends on a number of parameters such as income, wealth, family size, and retirement status [6]. It is assumed that the wealth variable is missed in the survey. We consider the following log model for household consumption [6], [69]:

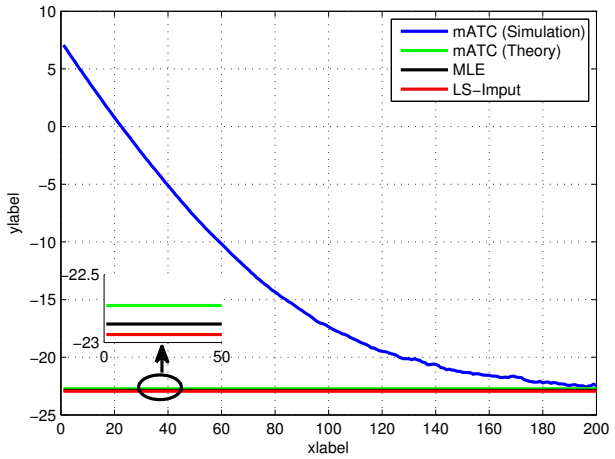
$$\begin{aligned}\ln \boldsymbol{c}_k(i) =& \alpha + (\ln \boldsymbol{l}_{k,i})\beta_1 + (\ln \boldsymbol{m}_{k,i}^{p})\beta_2 + (\ln \boldsymbol{m}_{k,i}^{c})\beta_3 \\ &+ \boldsymbol{t}_{k,i}\beta_4 + \boldsymbol{v}_k(i)\end{aligned} \quad (129)$$

where α is $\ln \boldsymbol{c}_k(i)$ intercept, $\boldsymbol{c}_k(i)$ is the consumption of household k at time i, $\boldsymbol{l}_{k,i}$ is the total wealth, which is assumed to be censored, $\boldsymbol{m}_{k,i}^{p}$ is the permanent part of the income (a long-term measurement of average income that depends on a number of parameters such as family income and education) [69]–[73], $\boldsymbol{t}_{k,i}$ refers to the retirement status and family size. The modeling error $\boldsymbol{\epsilon}_k(i)$ is assumed to be zero-mean. In a manner similar to [6], we only consider the first 4 components of the regressor, i.e., we set $\beta_4 = 0$. As suggested by the earlier Remark 1, we subtract the mean of the measurements, which is given by

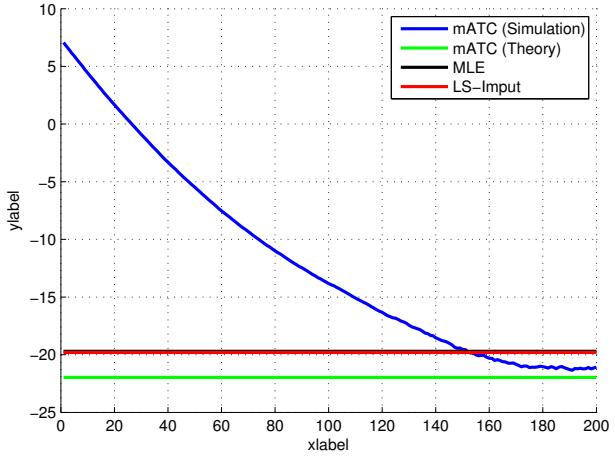
$$\begin{aligned}\mathbb{E}\,(\ln \boldsymbol{c}_k(i)) =& \alpha + \mathbb{E}(\ln \boldsymbol{l}_{k,i})\beta_1 + \mathbb{E}(\ln \boldsymbol{m}_{k,i}^{p})\beta_2 \\ &+ \mathbb{E}(\ln \boldsymbol{m}_{k,i}^{c})\beta_3 + \mathbb{E}\boldsymbol{t}_{k,i}\beta_4\end{aligned}$$

from both sides of (129) and arrive at the model

$$\boldsymbol{d}_k^c(i) = \boldsymbol{u}_{k,i}^c w_c + \boldsymbol{v}_k(i) \quad (130)$$



(a)



(b)

Fig. 2. The MSD learning curves for different diffusion algorithms at node 1 for (a) $p = 0.3$ and (b) $p = 0.4$.

where

$$\begin{aligned} w_c &= [\beta_1 \ \beta_2 \ \beta_3]^T \\ \boldsymbol{d}_k^c(i) &\triangleq \ln \boldsymbol{c}_k(i) - \mathbb{E}\,(\ln \boldsymbol{c}_k(i)) \\ \boldsymbol{u}_{k,i}^c &\triangleq [\ln \boldsymbol{l}_{k,i} \ \ln \boldsymbol{m}_{k,i}^p \ \ln \boldsymbol{m}_{k,i}^c] - \mathbb{E}\,[\ln \boldsymbol{l}_{k,i} \ \ln \boldsymbol{m}_{k,i}^p \ \ln \boldsymbol{m}_{k,i}^c]. \end{aligned} \tag{131}$$

Using a complete set of data, the authors in [6] reported the estimate $\widehat{w}_c = [0.054,\ 0.182,\ 0.204]^T$ for the unknown parameters. We generate data according to $\widehat{w}_c$ and assume that the regressor $\boldsymbol{u}_{k,i}^c$ has Gaussian distribution. We model $\boldsymbol{v}_k(i)$ by a zero-mean Gaussian random variable. We further assume that the log of wealth is randomly missed and we consider a uniform distribution over $[-\bar{q}, \bar{q}]$ for the missing variable, thus $\sigma_{\xi,k}^2 = \bar{q}^2/3$. In the simulation, we use $\bar{q} = 0.5$. In the survey, it has been observed that approximately 30% of total wealth, including housing and stock market, is censored [6], i.e., $p = 0.3$. In the simulation, we use $\mu = 0.025$. The updating step for estimating the variance $\sigma_{\xi,k}^2$ is executed from the beginning $i = 1$. We also set $\alpha_1 = \alpha_2 = 0.001$, and $\alpha_3 = 0.0001$.

To derive a decision rule for this example, we first consider

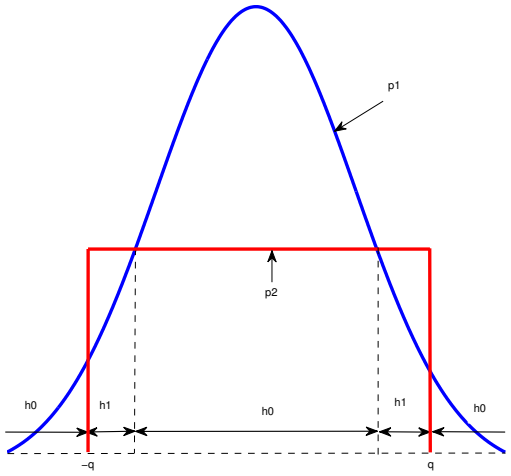


Fig. 3. Decision regions and scaled probabilities with prior beliefs about $\mathcal{H}_0$ and $\mathcal{H}_1$.

two distributions under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$:

$$p_{\mathcal{H}_0}(\bar{\boldsymbol{u}}_{k,i}^{cj}) = \frac{1}{\sqrt{2\pi R_{u^c,k}(j,j)}} \exp\left[-\frac{|\bar{\boldsymbol{u}}_{k,i}^{cj}|^2}{2R_{u^c,k}(j,j)}\right] \tag{132}$$

$$p_{\mathcal{H}_1}(\bar{\boldsymbol{u}}_{k,i}^{cj}) = \begin{cases} \frac{1}{2\bar{q}}, & \text{if } |\bar{\boldsymbol{u}}_{k,i}^{cj}| \le \bar{q} \\ 0, & \text{otherwise} \end{cases} \tag{133}$$

In the scenario for $|\bar{\boldsymbol{u}}_{k,i}^{j}| \le \bar{q}$ the LLR is obtained as

$$\text{LLR} = \frac{\sqrt{2\pi R_{u^c,k}(j,j)}}{2\bar{q}} \exp\left[\frac{|\bar{\boldsymbol{u}}_{k,i}^{cj}|^2}{2R_{u^c,k}(j,j)}\right], \quad |\bar{\boldsymbol{u}}_{k,i}^{j}| \le \bar{q}. \tag{134}$$

Since $\bar{q}$ and $R_{u^c,k}(j,j)$ are unknown *a priori*, we need to approximate them from the data. Similar to the previous case, we assume a good approximation of the ratio r_k is available that helps us to obtain estimates of $\bar{q}$ and $R_{u^c,k}$ similar to (127)–(128). Then, we find an estimate for $\widehat{\bar{q}}$ as

$$\begin{aligned} \widehat{\bar{q}} &= \sqrt{3\check{\sigma}_{\xi,k}^2} \\ &= \sqrt{\frac{3}{M_k\big((1-\widehat{p})r_k + \widehat{p}\big)} \sum_{i=1}^{M_k} \bar{\boldsymbol{u}}_{k,i}^{cj} \bar{\boldsymbol{u}}_{k,i}^{cj*}}. \end{aligned} \tag{135}$$

Using (134) and (121), we obtain the following decision rule:

$$\text{if } |\bar{\boldsymbol{u}}_{k,i}^{j}| \le \widehat{\bar{q}} : |\bar{\boldsymbol{u}}_{k,i}^{cj}|^2 \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{>}} 2\widehat{R}_{u^c,k}(j,j) \log\left(\frac{2(1-\widehat{p})\widehat{\bar{q}}}{\widehat{p}\sqrt{2\pi \widehat{R}_{u^c,k}(j,j)}}\right) \tag{136}$$

where we replaced $R_{u^c,k}(j,j)$ and $\bar{q}$, respectively, by their estimates $\widehat{R}_{u^c,k}(j,j)$ and $\widehat{\bar{q}}$. Figure 3 shows an example of two distributions scaled by $\widehat{p}$ and $(1-\widehat{p})$ and the corresponding decision regions.

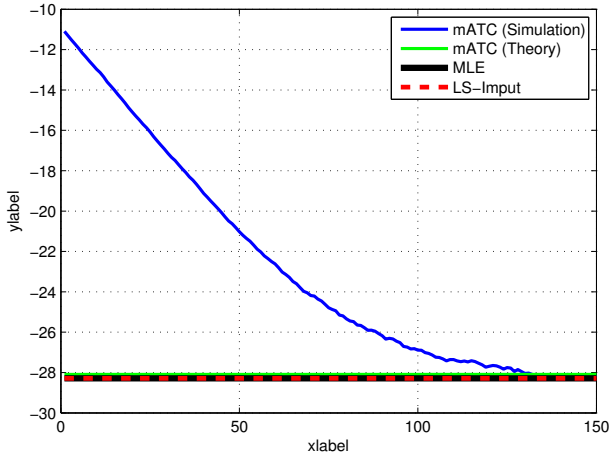


Fig. 4. The MSD of different algorithms for the household data for node 1.

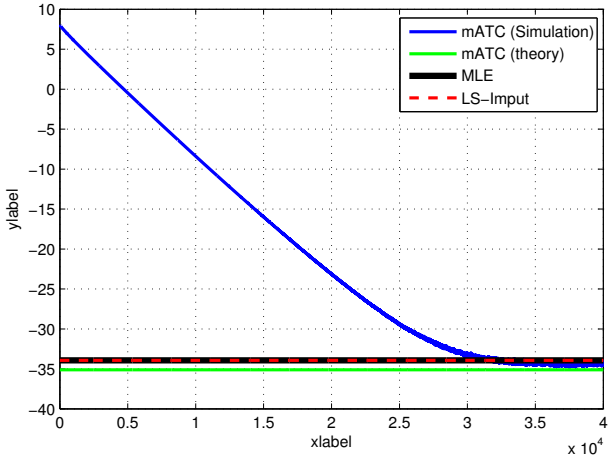


Fig. 5. The MSD of different approaches for the mental health survey data for node 1.

Figure 4 shows the MSDs of the estimators for the household consumption data. As it is observed, the proposed mATC shows comparable performance with the centralized approach.

B. Mental Health Survey

In the last simulation, we consider the following model, motivated by a mental health survey study run by various companies [65], [74]:

$$\bar{\boldsymbol{d}}_k(i) = \beta + \widetilde{\boldsymbol{u}}_{k,i} w + \boldsymbol{v}_k(i), \quad i = 1, 2, \ldots, N \tag{137}$$

where $\bar{\boldsymbol{d}}_k(i)$ is the square root of the total depression score for every individual i, β is the $\bar{\boldsymbol{d}}_k(i)$ intercept, $\widetilde{\boldsymbol{u}}_{k,i}$ denotes the regressor (covariate) for every individual i, and $\boldsymbol{v}_k(i)$ is the modeling error. Index k refers to the company index and i is used for participant's index. The elements of $\widetilde{\boldsymbol{u}}_{k,i}$, defined in [74] are shown in Table I and they include variables such as income, age, and martial status.

The square root of the total depression score $\bar{\boldsymbol{d}}_k(i)$ is measured based on individual answers to 20 questions regarding feeling about depression [74]. For example, the answer to the second question "I felt depressed" can be a number between 0 and 3; 0 for less than 1 day feeling, 1 for one to two days, 2 for occasionally or a moderate amount of time (3 to 4 days), and 3 for most of the time (5 to 7 days) [74].

We apply the least squares technique to a subset of the data provided in [74] to find an estimate for $[\widehat{\alpha}\ \widehat{w}^T]$ in (137) as $[\widehat{\alpha}\ \widehat{w}^T] = [0.1, 0.27, -0.03, -0.06, 0.13, 0.73, -0.28, 0.22]$. We then use the estimate for $\widehat{w}$ to generate zero-mean square root total depression scores as follows. Again, as indicated by the earlier Remark 1 and similar to the previous application, we modify the model of (137) as follows. Consider the mean of both sides of model (137), which is given by

$$\mathbb{E}\bar{\boldsymbol{d}}_k(i) = \beta_0 + (\mathbb{E}\widetilde{\boldsymbol{u}}_{k,i}) w, \tag{138}$$

then, we subtract the above mean from both sides (137) to get

$$\begin{aligned} \boldsymbol{d}_k(i) &\triangleq \bar{\boldsymbol{d}}_k(i) - \mathbb{E}\bar{\boldsymbol{d}}_k(i) \\ &= \boldsymbol{u}_{k,i} w + \boldsymbol{v}_k(i), \quad i = 1, 2, \ldots, N, \end{aligned} \tag{139}$$

where $\boldsymbol{u}_{k,i} = \widetilde{\boldsymbol{u}}_{k,i} - \mathbb{E}\widetilde{\boldsymbol{u}}_{k,i}$ is a zero-mean random vector. To generate $\boldsymbol{u}_{k,i}$, we uniformly generate the regressor $\tilde{\boldsymbol{u}}_{k,i}$ according to Table I and then subtract the mean. We further assume that the income is missed with probability 0.3 in the simulation study. We consider a zero-mean Gaussian distribution with variance 0.004 for missing parts, i.e., $\sigma^2_{\xi,k} = 0.004$.

The covariance matrix of the discrete regressor $\boldsymbol{u}_{k,i}$ is given by

$$R_{u,k} = \mathrm{diag}\{0.25, 252, 2, 2.967, 0.11, 1.25, 4\}. \tag{140}$$

The algorithm needs a smaller step size than the one in the previous simulation to converge due to the largest eigenvalue of the covariance matrix in (140). In the simulation, we set $\mu = 0.0025$. We assume a Gaussian distribution for measurement noise, $\boldsymbol{v}_k(i) \sim \mathcal{N}(0, \sigma^2_{v,k})$ with $\sigma^2_{v,k} = 0.01$. In this scenario, the updating for the estimate of variance $\widehat{\boldsymbol{\sigma}}^2_{\xi,k}(i)$ starts from the beginning $i = 1$. Hence, the proposed algorithm is expected to have a slower convergence rate at smaller step-sizes. We set $\alpha_1 = \alpha_2 = \alpha_3 = 0.0001$.

Figure 5 shows the MSDs of different approaches for mental health survey. As it is observed the proposed mATC approach shows promising results compared to centralized techniques. One way to improve the convergence rate is to estimate the curvature information of the cost function and modify the update step (37) based on the curvature information. Finally, Table II shows the estimate of the variance $\sigma^2_{\xi,k}$ at different nodes in steady state. It is seen that every node can obtain a good estimate for $\sigma^2_{\xi,k}$ using the proposed mATC algorithm.

VI. CONCLUSIONS

In this paper, we examined the estimation of an unknown vector over a connected network of agents, with each agent subjected to a stream of data with incomplete regressors. We have shown that the estimator in general is biased; hence, we have modified the cost function by a (de)regularisation term to mitigate the bias and obtained a distributed approach based on diffusion adaptation techniques. We have also suggested a technique to estimate the (de)regularization term from the

TABLE I
DEPRESSION COVARIATE FOR THE i-TH INDIVIDUAL AT THE $k-$TH COMPANY $\left(\widetilde{\boldsymbol{u}}_{k,i} = [\widetilde{\boldsymbol{u}}^1_{k,i}, \widetilde{\boldsymbol{u}}^2_{k,i}, \widetilde{\boldsymbol{u}}^3_{k,i}, \widetilde{\boldsymbol{u}}^4_{k,i}, \widetilde{\boldsymbol{u}}^5_{k,i}, \widetilde{\boldsymbol{u}}^6_{k,i}, \widetilde{\boldsymbol{u}}^7_{k,i}]\right)$ [74].

Covariate	Range
Gender: $\widetilde{\boldsymbol{u}}^1_{k,i}$	=1 if the i-th responder is male; =2 for the female
Age in year: $\widetilde{\boldsymbol{u}}^2_{k,i}$	19,20,...,73
Marital: $\widetilde{\boldsymbol{u}}^3_{k,i}$	=1 never married; =2 married; =3 divorced;=4 separated; =5 widowed
Education: $\widetilde{\boldsymbol{u}}^4_{k,i}$	=1 less than high school; =2 some high school; =3 finished high school; =4 some college; =5 finished bachelor's degree; =6 finished master's degree; =7 finished doctorate
Log of Income: $\widetilde{\boldsymbol{u}}^5_{k,i}$	$\log 4000$ to $\log 55000$
Religion: $\widetilde{\boldsymbol{u}}^6_{k,i}$	=1 Protestant; =2 Catholic; =3 Jewish; =4 none; =5 other
Employment: $\widetilde{\boldsymbol{u}}^7_{k,i}$	=1 full time; =2 part time; =3 unemployed; =4 retired; =5 houseperson; =6 in school; =7 other

TABLE II
ESTIMATES OF VARIANCE $\sigma^2_{\xi,k}$ IN DIFFERENT NODES FOR THE MENTAL HEALTH SURVEY DATA. THE TRUE VALUE OF THE VARIANCE IS $\sigma^2_{\xi,k} = 0.004$.

Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7
0.0047	0.0046	0.0048	0.0049	0.0046	0.0049	0.0047

data. We have studied the performance of the proposed algorithm under some simplifying assumptions and considered two applications in mental health and household consumption surveys. Simulation results show a comparable performance compared to existing centralized approaches based on imputation techniques.

REFERENCES

[1] M. R. Gholami, E. G. Ström, and A. H. Sayed, "Distributed estimation over cooperative networks with missing data," in *Proc. IEEE GlobalSIP*, Austin, TX, Dec. 2013, pp. 411–414.

[2] C. K. Enders, *Applied Missing Data Analysis*. The Guilford Press, 2010.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, "Missing-data methods for generalized linear models: A comparative review," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, 2005.

[5] P. W. Bernhardt, H. J. Wang, and D. Zhang, "Statistical methods for generalized linear models with covariates subject to detection limits," *Statistics in Biosciences*, pp. 1–22, 2013.

[6] R. Rigobon and T. M. Stoker, "Estimation with censored regressors: Basic issues," *International Economic Review*, vol. 48, no. 4, pp. 1441–1467, 2007.

[7] D. R. Cox and D. Oakes, *Analysis of Survival Data*. Chapman & Hall/CRC, 1984.

[8] C. K. Enders, "A primer on the use of modern missing-data methods in psychosomatic medicine research," *Psychosomatic Medicine*, vol. 68, no. 3, pp. 427–436, 2006.

[9] B. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, "Collaborative filtering and the missing at random assumption," *arXiv:1206.5267*, 2012.

[10] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5–37, 2010.

[11] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[12] A. Davey and et al., *Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach*. Routledge Academic, 2009.

[13] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley, 2002.

[14] R. Rigobon and T. M. Stoker, "Bias from censored regressors," *Journal of Business & Economic Statistics*, vol. 27, no. 3, pp. 340–353, 2009.

[15] P.-L. Loh and M. J. Wainwright, "Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2012, pp. 2601–2605, Cambridge, MA.

[16] G. Arminger, C. C. Clogg, and M. E. Sobel, *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Springer, 1995.

[17] C. Wang and M. C. Paik, "Efficiencies of methods dealing with missing covariates in regression analysis," *Statistica Sinica*, vol. 16, no. 4, p. 1169, 2006.

[18] H. J. Wang and X. Feng, "Multiple imputation for M-regression with censored covariates," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 194–204, 2012.

[19] P. Royston, "Multiple imputation of missing values," *Stata Journal*, vol. 4, pp. 227–241, 2004.

[20] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 2009, vol. 307.

[21] N. J. Horton and N. M. Laird, "Maximum likelihood analysis of generalized linear models with missing covariates," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 37–50, 1999.

[22] I. M. Trawinski and R. Bargmann, "Maximum likelihood estimation with incomplete multivariate data," *The Annals of Mathematical Statistics*, vol. 35, no. 2, pp. 647–657, 1964.

[23] A. Wiesel, Y. C. Eldar, and A. Yeredor, "Linear regression with Gaussian model uncertainty: Algorithms and bounds," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2194–2205, 2008.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[25] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," in *International Conference on Image Processing,*, vol. 1, 2001, pp. 145–148.

[26] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, 1994, pp. 120–127.

[27] L. Li, C. Shen, X. Li, and J. M. Robins, "On weighting approaches for missing data," *Statistical methods in medical research*, pp. 14–30, 2011.

[28] J. V. Tsimikas, L. E. Bantis, and S. D. Georgiou, "Inference in generalized linear regression models with a censored covariate," *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1854–1868, 2012.

[29] W. Nelson and G. J. Hahn, "Linear estimation of a regression relationship from censored data part I- simple methods and their application," *Technometrics*, vol. 14, no. 2, pp. 247–269, 1972.

[30] ——, "Linear estimation of a regression relationship from censored data-part II best linear unbiased estimation and theory," *Technometrics*, vol. 15, no. 1, pp. 133–150, 1973.

[31] D. Thanou, E. Kokiopoulou, Y. Pu, and P. Frossard, "Distributed average consensus with quantization refinement," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 194–205, Jan. 2013.

[32] A. Dogandzic and K. Qiu, "Decentralized random-field estimation for sensor networks using quantized spatially correlated data and fusion-center feedback," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 6069–6085, 2008.

[33] J. Fang and H. Li, "Optimal/near-optimal dimensionality reduction for distributed estimation in homogeneous and certain inhomogeneous scenarios," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4339–4353, 2010.

[34] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, pp. 311–801, issue 4–5, NOW Publishers, July 2014.

[35] ——, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, pp. 460–497, Apr. 2014.

[36] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. Towfic, "Diffusion strategies for adaptation and learning," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 155–171, May 2013.

[37] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, vol. 3, R. Chellapa and S. Theodoridis, Eds., pp. 323–454, Academic Press, Elsevier, 2014.

[38] J. Chen and A. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[39] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[40] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1867, Nov. 2010.

[41] P. D. Lorenzo and S. Barbarossa, "A bio-inspired swarming algorithm for decentralized access in cognitive radio," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6160–6174, Dec. 2011.

[42] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *Proc. Int. Conf. Inform. Fusion (FUSION)*, Cologne, Germany, Jun. - Jul. 2008, pp. 1–6.

[43] J. Fang and H. Li, "Distributed consensus with quantized data via sequence averaging," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 944–948, 2010.

[44] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, pp. 6217–6234, Dec. 2012.

[45] B. Efron, "Censored data and the bootstrap," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 312–319, 1981.

[46] B. W. Turnbull, "The empirical distribution function with arbitrarily grouped, censored and truncated data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 38, no. 3, pp. 290–295, 1976.

[47] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[48] V. Chernozhukov, R. Rigobon, and T. M. Stoker, "Set identification with tobin regressors," cemmap working paper, Tech. Rep., 2009.

[49] D. Hedeker and R. D. Gibbons, "Application of random-effects pattern-mixture models for missing data in longitudinal studies," *Psychological methods*, vol. 2, no. 1, p. 64, 1997.

[50] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," *Intell. Sys. Mol. Biol.*, 1994.

[51] R. J. Little, "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 125–134, 1993.

[52] C. E. Rasmussen, "The infinite gaussian mixture model," in *NIPS*, vol. 12, 1999, pp. 554–560.

[53] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through bayesian sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 363–375, 1994.

[54] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall New Jersey, 2000, vol. 1.

[55] A. H. Sayed, *Adaptive Filters*. New York: Wiley, 2008.

[56] R. Abdolee, B. Champagne, and A. H. Sayed, "A diffusion LMS strategy for parameter estimation in noisy regressor applications," in *Proc. EUSIPCO*, Bucharest, Romania, August 2012, pp. 749–753.

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

[58] G. Casella and R. L. Berger, *Statistical Inference*. 2nd ed., Duxbury, Pacific Grove, CA, USA, 2002.

[59] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[60] R. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2245–2253, Aug. 2003.

[61] D. Gu, "Distributed EM algorithm for Gaussian mixtures in sensor networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 7, pp. 1154–1166, Jul. 2008.

[62] C. Plant and C. Bohm, "Parallel EM-clustering: Fast convergence by asynchronous model updates," in *IEEE International Conference on Data Mining Workshops (ICDMW)*, Dec. 2010, pp. 178–185.

[63] H. V. Poor, *An introduction to signal detection and estimation*. Springer, 1994.

[64] S. M. Kay, "Fundamentals of statistical signal processing: Detection theory," 1998.

[65] R. J. Little, "Regression with missing x's: A review," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1227–1237, 1992.

[66] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[67] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[68] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, 2001.

[69] J. A. Parker, "Spendthrift in America? On two decades of decline in the US saving rate," in *NBER Macroeconomics Annual 1999*, 2000, vol. 14, pp. 317–387.

[70] J. Y. Campbell and N. G. Mankiw, "Permanent income, current income, and consumption," *Journal of Business & Economic Statistics*, vol. 8, no. 3, pp. 265–279, 1990.

[71] A. Lusardi, "Permanent income, current income, and consumption: Evidence from two panel data sets," *Journal of Business & Economic Statistics*, vol. 14, no. 1, pp. 81–90, 1996.

[72] R. E. Hall, "Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence," *NBER Working Paper*, no. R0015, 1979.

[73] M. Friedman, *A Theory of the Consumption Function*. Princeton University Press, 1957.

[74] A. A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*. CRC Press, 2004.

Mohammad Reza Gholami (S'09–M'14) received the Ph.D. degree from Chalmers University of Technology, Gothenburg, Sweden, in 2013. From 2012 to 2013, he was a visiting researcher with the Adaptive Systems Laboratory, University of California at Los Angeles. From 2014 to 2015, he was a Post-Doctoral Research Fellow with the Department of Signal Processing, KTH Royal Institute of Technology, Stockholm, Sweden. He is currently a Senior Data Scientist with Campanja AB. His main interests are statistical inference, distributed estimation, wireless network positioning, network wide synchronization, and data analytics.

Magnus Jansson received the Master of Science, Technical Licentiate, and Ph.D. degrees in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden, in 1992, 1995 and 1997, respectively. During 1997-98 he had a lecturer position in the Control Department at KTH. Since 1998 he has had various positions in the Signal Processing Department at KTH: 1998-2003 Assistant Professor, 2003-2012 Associate Professor, and since 2013 Professor. In January 2002 he was appointed Docent in Signal Processing at KTH. Dr Jansson spent one year during 1998-99 as a visiting researcher at the Department of Electrical and Computer Engineering, University of Minnesota. Dr Jansson served as Associate Editor of IEEE Signal Processing Letters 2008-2012, Senior Area Editor of IEEE Signal Processing Letters (2012-2014). Currently he is Associate Editor of EURASIP Journal on Advances in Signal Processing (2007-), and EURASIP Signal Processing (2015-). His research interests include statistical signal processing; navigation and positioning, sensor array processing, time series analysis, and system identification.

Erik G. Ström (S'93–M'95–SM'01) received the M.S. degree from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1990, and the Ph.D. degree from the University of Florida, Gainesville, in 1994, both in electrical engineering. He accepted a postdoctoral position at the Department of Signals, Sensors, and Systems at KTH in 1995. In February 1996, he was appointed Assistant Professor at KTH, and in June 1996 he joined Chalmers University of Technology, Göteborg, Sweden, where he is now a Professor in Communication Systems since June 2003. Dr. Ström currently heads the Division for Communications Systems and leads the competence area Sensors and Communications at the traffic safety center SAFER, which is hosted by Chalmers. His research interests include signal processing and communication theory in general, and constellation labelings, channel estimation, synchronization, multiple access, medium access, multiuser detection, wireless positioning, and vehicular communications in particular. Since 1990, he has acted as a consultant for the Educational Group for Individual Development, Stockholm, Sweden. He is a contributing author and associate editor for Roy. Admiralty Publishers FesGas-series, and was a co-guest editor for the *Proceedings of the IEEE* special issue on Vehicular Communications (2011) and the *IEEE Journal on Selected Areas in Communications* special issues on Signal Synchronization in Digital Transmission Systems (2001) and on Multiuser Detection for Advanced Communication Systems and Networks (2008). Dr. Ström was a member of the board of the IEEE VT/COM Swedish Chapter 2000–2006. He received the Chalmers Pedagogical Prize in 1998 and the Chalmers Ph.D. Supervisor of the Year award in 2009.

A. H. Sayed is distinguished professor and former chairman of electrical engineering at UCLA where he directs the UCLA Adaptive Systems Laboratory (www.ee.ucla.edu/asl). An author of over 480 scholarly publications and six books, his research involves several areas including adaptation and learning, statistical signal processing, network science, information processing theories, and biologically-inspired designs. His work has been recognized with several awards including the 2015 Education Award from the IEEE Signal Processing Society, the 2014 Papoulis Award from the European Association for Signal Processing, the 2013 Meritorious Service Award and the 2012 Technical Achievement Award from the IEEE Signal Processing Society, the 2005 Terman Award from the American Society for Engineering Education, the 2003 Kuwait Prize, and the 1996 IEEE Donald G. Fink Prize. He has been awarded several Best Paper Awards from the IEEE (2002, 2005, 2012, 2014) and EURASIP (2015) and is a Fellow of both the IEEE and the American Association for the Advancement of Science (AAAS). He is recognized as a Highly Cited Researcher by Thomson Reuters. He is President-Elect of the IEEE Signal Processing Society (2016-2017).