LOW-RANK DATA MATRIX RECOVERY WITH MISSING VALUES AND FAULTY SENSORS

Roberto López-Valcarce[*]

Universidade de Vigo, Spain

Josep Sala-Alvarez[†]

Universitat Politècnica de Catalunya, Spain

ABSTRACT

In practice, data gathered by wireless sensor networks often belongs in a low-dimensional subspace, but it can present missing as well as corrupted values due to sensor malfunctioning and/or malicious attacks. We study the problem of Maximum Likelihood estimation of the low-rank factors of the underlying structure in such situation, and develop an Expectation-Maximization algorithm to this purpose, together with an effective initialization scheme. The proposed method outperforms previous schemes based on an initial faulty sensor identification stage, and is competitive in terms of complexity and performance with convex optimization-based matrix completion approaches.

***Index Terms*—** Low-rank approximation, matrix completion, outliers, faulty sensors, wireless sensor networks.

1. INTRODUCTION

The fact that many high-dimensional data structures can be accurately represented as lying in a low-dimensional subspace has spurred a number of approaches to exploit such low-rank structure [1]. In particular, in wireless sensor networks (WSNs), the measurements taken by K sensors over N sensing slots can be organized into a $K \times N$ data matrix, which in many practical applications will be close to having rank $r \ll \min\{K, N\}$. This is because environmental data usually exhibits strong correlation across space and time [2]. When applied to WSNs, matrix completion techniques [1, 3], which exploit this property in order to recover the whole data matrix from a subset of its samples, may result in significant savings in sensing and communication [4, 5]. Missing data may also occur if a transmission from a sensor to the Fusion Center (FC) is affected by a deep channel fade.

WSNs consist of low-cost, battery-powered devices often operating in harsh environments, and therefore prone to sensor malfunction [6, 7], making data generated by faulty sensors unreliable. Additionally, sensors are usually unattended, which makes them vulnerable to data manipulation by malicious external agents [8,9]. Regardless of its origin, corrupted samples may significantly degrade data analysis at the FC.

In this paper we investigate algorithms that estimate the low-rank factors of the underlying data matrix in the presence of both missing data and unreliable sensors. In particular, we adopt a non-informative model under which faulty sensors, whose number and identities are not known, produce i.i.d. data samples uniformly distributed in some unknown interval. In this setting, Maximum Likelihood (ML) estimators are attractive because of their good asymptotic performance properties [15]. The presence of hidden variables (i.e., missing data, identities of faulty sensors) precludes a closed-form solution for the ML estimators and motivates the application of the Expectation-Maximization (E-M) algorithm [16], which iteratively refines the estimates of the low-rank factors as well as *a posteriori* probabilities of individual sensor faults.

Matrix completion methods have become increasingly popular in applications in which a low-rank matrix is to be recovered from a limited number of observations. Typically, they seek to minimize the nuclear norm (which is a convex surrogate of the rank) under some constraint on the fidelity of the reconstruction to account for missing data and/or measurement noise [17,18]. When the underlying rank is assumed known, non-convex approaches based on the so-called Burer-Monteiro factorization [19] are also available [20, 21]. In the presence of corrupted data, convex minimization techniques have also been proposed by exploiting the fact that such anomalous data can be assumed to be sparse [22–24]. Many of these methods enjoy theoretical performance guarantees under suitable conditions [1, 3]; nevertheless, they invariably include regularization terms whose parameters need appropriate tuning. This is not the case with the proposed E-M approach, which iteratively estimates all unknown parameters; in addition, simulation results show that E-M may outperform sparsity-based methods even with optimal tuning.

2. PROBLEM STATEMENT

Consider a network of K devices collecting data from their environment. Each device, say sensor i, gathers N data points $\{y_{ij},\, j = 1, \ldots, N\}$, to be collected at the FC in the $K \times N$ data matrix $\boldsymbol{Y}$. It is assumed that the underlying physical phenomenon gives rise to a matrix $\boldsymbol{LR}$ having low rank $r \ll$

[*] Supported by Agencia Estatal de Investigación (Spain) and the European Regional Development Fund (ERDF) under project WINTER (TEC2016-76409-C2-2-R), and by Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019).

[†] Supported by Agencia Estatal de Investigación (Spain) and ERDF under project WINTER (TEC2016-76409-C2-1-R), and the Catalan administration (AGAUR) under 2017 SGR 578.

$\min\{K, N\}$, with $\boldsymbol{L} \in \mathbb{R}^{K\times r}$ and $\boldsymbol{R} \in \mathbb{R}^{r\times N}$. Measurements are noisy, so that under ideal conditions the data matrix would be given by $\boldsymbol{LR}+\boldsymbol{N}$, where $\boldsymbol{N}$ is a noise matrix with i.i.d. zero-mean Gaussian entries with variance σ^2. However, a number of sensors may be malfunctioning, in which case their collected data do not adhere to the above model. Thus, let $a_i = 0$ if sensor i is faulty, and $a_i = 1$ otherwise, and collect these in $\boldsymbol{a} = \left[\begin{array}{ccc} a_1 & \cdots & a_K \end{array}\right]^T$. Whether a given device is faulty or not is unknown at the FC. We assume that sensor faults take place randomly and independently, with the same *a priori* probability q for all devices, i.e., the a_i's are i.i.d. Bernoulli random variables with $\Pr\{a_i = 0\} = q$. We adopt a non-informative model under which the readings from a defective device, say sensor i, are i.i.d. and uniformly distributed in some unknown interval $[A_i, B_i]$.

Due to channel fading, device battery outages, etc., a subset of the observations is missing at the FC. We denote by Ω_i the set of pairs (i, j) for which y_{ij} is available, with cardinality $|\Omega_i| = N_i$, and define $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$, with $|\Omega| = \sum_{i=1}^K N_i \triangleq M$. Then, arranging the available data in the vector $\boldsymbol{y} \in \mathbb{R}^M$, the data model can be expressed as

$$\boldsymbol{y} = \mathcal{P}_\Omega\left(\boldsymbol{A}(\boldsymbol{LR}+\boldsymbol{N}) + (\boldsymbol{I}-\boldsymbol{A})\boldsymbol{W}\right), \tag{1}$$

where $\mathcal{P}_\Omega(\boldsymbol{X})$ extracts the entries of $\boldsymbol{X}$ with indices in Ω, $\boldsymbol{A} \triangleq \mathrm{diag}\{\boldsymbol{a}\}$, and $\boldsymbol{W}$ is a random matrix whose i-th row entries are i.i.d., uniformly distributed in $[A_i, B_i]$.

Our goal is to estimate the low-rank factors $\boldsymbol{L}$, $\boldsymbol{R}$ given $\boldsymbol{y}$ and Ω. If all sensors were working properly, and in the absence of missing data, the ML estimates should minimize $\|\boldsymbol{Y} - \boldsymbol{LR}\|_F^2$, where $\boldsymbol{Y} = \mathrm{unvec}(\boldsymbol{y})$. By the Eckart-Young theorem [10], these can be readily obtained from the SVD of $\boldsymbol{Y}$ as $\hat{\boldsymbol{L}} = \boldsymbol{UB}^{-1}$ and $\hat{\boldsymbol{R}} = \boldsymbol{BSV}^T$, where $\boldsymbol{S}$ is diagonal with the r largest singular values, the columns of $\boldsymbol{U}$, $\boldsymbol{V}$ comprise the corresponding left and right singular vectors, and $\boldsymbol{B}$ is $r \times r$ invertible but otherwise arbitrary. However, with (unknown) faulty sensors and missing data, the ML estimates cannot be obtained in closed form.

3. E-M ALGORITHM DERIVATION

The E-M algorithm provides a computationally efficient means to iteratively seek a maximum of the likelihood function in the presence of incomplete observations [16]. In this context, we regard $\boldsymbol{y}$ as the *incomplete dataset,* and $\boldsymbol{z} = \{\boldsymbol{T}, \boldsymbol{a}\}$ as the *complete dataset,* with $\boldsymbol{T} \triangleq \boldsymbol{A}(\boldsymbol{LR}+\boldsymbol{N}) + (\boldsymbol{I}-\boldsymbol{A})\boldsymbol{W}$. The unknown parameters are collected in $\boldsymbol{\theta} = \{\boldsymbol{L}, \boldsymbol{R}, \sigma^2, q, A_1, \ldots, A_K, B_1, \ldots, B_K\}$.

The general form of the E-M algorithm is as follows. Given the incomplete dataset $\boldsymbol{y}$ and an initial guess $\hat{\boldsymbol{\theta}}_0$, at iteration k one performs the following:

- *E-step:* Compute the conditional expectation $Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) \triangleq \mathbb{E}_{\boldsymbol{z}|\boldsymbol{y}}\{\ln p(\boldsymbol{z}; \boldsymbol{\theta}) \,|\, \boldsymbol{y}\,;\, \hat{\boldsymbol{\theta}}_k\}$, where $p(\boldsymbol{z}; \boldsymbol{\theta})$ is the pdf of $\boldsymbol{z}$ parameterized by $\boldsymbol{\theta}$.
- *M-step:* the estimate is updated by maximizing this conditional expectation, i.e., $\hat{\boldsymbol{\theta}}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k)$.

3.1. Expectation step

Let $\boldsymbol{t}_i$ be the i-th column of $\boldsymbol{T}$. Due to independence, and using the Bernoulli pdf $p(a_i; \boldsymbol{\theta}) = q^{1-a_i}(1-q)^{a_i}$, one has

$$\begin{aligned} p(\boldsymbol{z}; \boldsymbol{\theta}) &= \prod_{i=1}^K p(\boldsymbol{t}_i, a_i; \boldsymbol{\theta}) = \prod_{i=1}^K p(\boldsymbol{t}_i \,|\, a_i; \boldsymbol{\theta}) p(a_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^K [q p_0(\boldsymbol{t}_i; \boldsymbol{\theta})]^{1-a_i} [(1-q) p_1(\boldsymbol{t}_i; \boldsymbol{\theta})]^{a_i}, \end{aligned} \tag{2}$$

where for convenience we denote $p_b(\boldsymbol{t}_i; \boldsymbol{\theta}) \triangleq p(\boldsymbol{t}_i | a_i = b; \boldsymbol{\theta})$, $b \in \{0, 1\}$. Now let us rewrite $Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k)$ as

$$\begin{aligned} Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) &= \mathbb{E}_{\boldsymbol{T},\boldsymbol{a}|\boldsymbol{y}}\{\ln p(\boldsymbol{z}; \boldsymbol{\theta}) \,|\, \boldsymbol{y}; \hat{\boldsymbol{\theta}}_k\} \\ &= \mathbb{E}_{\boldsymbol{a}|\boldsymbol{y}}\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{a},\boldsymbol{y}}\{\ln p(\boldsymbol{z}; \boldsymbol{\theta}) \,|\, \boldsymbol{a}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}\} \,|\, \boldsymbol{y}; \hat{\boldsymbol{\theta}}_k\right\}. \end{aligned} \tag{3}$$

Using (2), the inner conditional expectation in (3) becomes

$$\begin{aligned} \mathbb{E}_{\boldsymbol{T}|\boldsymbol{a},\boldsymbol{y}}\{\ln p(\boldsymbol{z}; \boldsymbol{\theta}) \,|\, \boldsymbol{a}, \boldsymbol{y}; \hat{\boldsymbol{\theta}}_k\} &= K \ln q + \left(\sum_{i=1}^K a_i\right) \ln \frac{1-q}{q} \\ &+ \sum_{i=1}^K (1-a_i) \mathbb{E}_{\boldsymbol{t}_i|a_i=0,\boldsymbol{y}}\{\ln p_0(\boldsymbol{t}_i; \boldsymbol{\theta}) \,|\, a_i = 0, \boldsymbol{y}\,;\, \hat{\boldsymbol{\theta}}_k\} \\ &+ \sum_{i=1}^K a_i \mathbb{E}_{\boldsymbol{t}_i|a_i=1,\boldsymbol{y}}\{\ln p_1(\boldsymbol{t}_i; \boldsymbol{\theta}) \,|\, a_i = 1, \boldsymbol{y}\,;\, \hat{\boldsymbol{\theta}}_k\}, \end{aligned} \tag{4}$$

since for any binary r.v. $a \in \{0, 1\}$, one has $a\mathbb{E}_{\boldsymbol{t}|a}\{f(\boldsymbol{t})\} = a[(1-a)\mathbb{E}_{\boldsymbol{t}|a=0}\{f(\boldsymbol{t})\} + a\mathbb{E}_{\boldsymbol{t}|a=1}\{f(\boldsymbol{t})\}] = a\mathbb{E}_{\boldsymbol{t}|a=1}\{f(\boldsymbol{t})\}$; analogously, $(1-a)\mathbb{E}_{\boldsymbol{t}|a}\{f(\boldsymbol{t})\} = (1-a)\mathbb{E}_{\boldsymbol{t}|a=0}\{f(\boldsymbol{t})\}$. Now, the entries t_{ij} of $\boldsymbol{t}_i$ are independent under both $a_i = 0$ and $a_i = 1$, so that $p_b(\boldsymbol{t}_i; \boldsymbol{\theta}) = \prod_{j=1}^N p_b(t_{ij}; \boldsymbol{\theta})$, $b \in \{0, 1\}$. Then, for $b \in \{0, 1\}$,

$$\begin{aligned} g_i^b(\boldsymbol{y}; \boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) &\triangleq \mathbb{E}_{\boldsymbol{t}_i|a_i=b,\boldsymbol{y}}\{\ln p_b(\boldsymbol{t}_i; \boldsymbol{\theta}) \,|\, a_i = b, \boldsymbol{y}\,;\, \hat{\boldsymbol{\theta}}_k\} \\ &= \sum_{j=1}^N \mathbb{E}_{t_{ij}|a_i=b,\boldsymbol{y}}\{\ln p_b(t_{ij}; \boldsymbol{\theta}) \,|\, a_i = b, \boldsymbol{y}\,;\, \hat{\boldsymbol{\theta}}_k\} \qquad (5) \\ &= \sum_{j\in\Omega_i} \ln p_b(y_{ij}; \boldsymbol{\theta}) + \sum_{j\notin\Omega_i} f_{ij}^b(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k), \qquad (6) \end{aligned}$$

where $f_{ij}^b(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) \triangleq \mathbb{E}_{t_{ij}|a_i=b}\{\ln p_b(t_{ij}; \boldsymbol{\theta}) \,|\, a_i = b; \hat{\boldsymbol{\theta}}_k\}$. Under our model, $p_0(t_{ij}; \boldsymbol{\theta}) = \mathcal{U}(t_{ij}; A_i, B_i)$ is a uniform pdf over $[A_i, B_i]$, whereas $p_1(t_{ij}; \boldsymbol{\theta}) = \mathcal{N}(t_{ij}; m_{ij}, \sigma^2)$ is a Gaussian pdf with mean $m_{ij} \triangleq (\boldsymbol{LR})_{ij}$ and variance σ^2. Hence, upon defining $\hat{m}_{ij,k} \triangleq (\hat{\boldsymbol{L}}_k \hat{\boldsymbol{R}}_k)_{ij}$, it follows that

$$f_{ij}^0(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) = \begin{cases} -\ln(B_i - A_i), & A_i \le \hat{A}_{i,k},\, B_i \ge \hat{B}_{i,k}, \\ -\infty, & \text{else}, \end{cases} \tag{7}$$

$$f_{ij}^1(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_k) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}[\hat{\sigma}_k^2 + (\hat{m}_{ij,k} - m_{ij})^2]. \tag{8}$$

Thus, letting $\hat{a}_{i,k} \triangleq \mathbb{E}_{a_i|\boldsymbol{y}}\{a_i|\boldsymbol{y};\hat{\boldsymbol{\theta}}_k\} = \Pr\{a_i = 1|\boldsymbol{y};\hat{\boldsymbol{\theta}}_k\}$,

$$Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k) \;=\; K\ln q + \left(\sum_{i=1}^{K}\hat{a}_{i,k}\right)\ln\frac{1-q}{q} + \sum_{i=1}^{K}\left[(1-\hat{a}_{i,k})g_i^0(\boldsymbol{y};\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k) + \hat{a}_{i,k}g_i^1(\boldsymbol{y};\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)\right]. \quad (9)$$

Note that $\hat{a}_{i,k}$ can be interpreted as the *a posteriori* probability of sesor i being non-defective, given the data $\boldsymbol{y}$ and the estimate $\hat{\boldsymbol{\theta}}_k$. Using Bayes' theorem, it is obtained as

$$\hat{a}_{i,k} = \frac{(1-\hat{q}_k)\prod_{j\in\Omega_i} p_1(y_{ij};\hat{\boldsymbol{\theta}}_k)}{\hat{q}_k\prod_{j\in\Omega_i} p_0(y_{ij};\hat{\boldsymbol{\theta}}_k) + (1-\hat{q}_k)\prod_{j\in\Omega_i} p_1(y_{ij};\hat{\boldsymbol{\theta}}_k)}. \quad (10)$$

3.2. Maximization step

In order to maximize $Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)$ in (9) w.r.t. $\boldsymbol{\theta}$, we proceed step by step. First, the optimum value of q is readily found:

$$\hat{q}_{k+1} = 1 - \frac{1}{K}\sum_{i=1}^{K}\hat{a}_{i,k}. \quad (11)$$

Second, note that the values of A_i, B_i maximizing $Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)$ are those maximizing $g_i^0(\boldsymbol{y};\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)$. This function evaluates to $-\infty$ unless $A_i \le \hat{A}_{i,k}$, $B_i \ge \hat{B}_{i,k}$, and $A_i \le y_{ij}$, $B_i \ge y_{ij}$ $\forall j \in \Omega_i$, in which case it evaluates to $-N\ln(B_i - A_i)$. Subject to the above constraints, this is maximized at $\hat{A}_{i,k+1} = \min\{\hat{A}_{i,k}\}\cup\{y_{ij}, j\in\Omega_i\}$, $\hat{B}_{i,k+1} = \max\{\hat{B}_{i,k}\}\cup\{y_{ij}, j\in\Omega_i\}$. Therefore, it is clear that we can simply take, for all k,

$$\hat{A}_{i,k} = \min_{j\in\Omega_i}\{y_{ij}\}, \qquad \hat{B}_{i,k} = \max_{j\in\Omega_i}\{y_{ij}\}. \quad (12)$$

Third, letting $c_i \triangleq 1 - \frac{N_i}{N}$, the values of $\boldsymbol{L}$, $\boldsymbol{R}$ maximizing $Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)$ in (9) are seen to be those minimizing the term

$$-2\sum_{i=1}^{K}\hat{a}_{i,k}g_i^1(\boldsymbol{y};\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k) \;=\; N\sum_{i=1}^{K}\hat{a}_{i,k}\left[\ln 2\pi\sigma^2 + c_i\frac{\hat{\sigma}_k^2}{\sigma^2}\right] + \frac{1}{\sigma^2}\sum_{i=1}^{K}\hat{a}_{i,k}\left[\sum_{j\in\Omega_i}(y_{ij}-m_{ij})^2 + \sum_{j\notin\Omega_i}(\hat{m}_{ij,k}-m_{ij})^2\right] \quad (13)$$

Only the last term in the right-hand side of (13) depends on $\boldsymbol{L}$, $\boldsymbol{R}$ via $\{m_{ij}\}$. If we define $\hat{\boldsymbol{Y}}_k \in \mathbb{R}^{K\times N}$ with entries

$$\hat{y}_{ij,k} = y_{ij}, \quad j\in\Omega_i, \qquad \hat{y}_{ij,k} = \hat{m}_{ij,k}, \quad j\notin\Omega_i, \quad (14)$$

and the diagonal matrix $\hat{\boldsymbol{A}}_k = \mathrm{diag}\{\ \hat{a}_{1,k}\ \cdots\ \hat{a}_{K,k}\ \}$, then such term can be rewritten so as to yield

$$\min_{\boldsymbol{L},\boldsymbol{R}}\frac{1}{\sigma^2}\left\|\hat{\boldsymbol{A}}_k^{1/2}(\hat{\boldsymbol{Y}}_k - \boldsymbol{L}\boldsymbol{R})\right\|_F^2, \quad (15)$$

which is a particular case of the *weighted low-rank approximation (WLRA) problem* [12]. Although general WLRA problems do not generally admit closed-form solutions, the special structure of (15), in which the weighting is on a row-by-row basis, constitutes an exception, as shown in [13]. This can be seen by noticing in (15) that $\hat{\boldsymbol{A}}_k^{1/2}\boldsymbol{L}$ and $\boldsymbol{R}$ are the factors in the (unweighted) rank-r approximation of $\hat{\boldsymbol{A}}_k^{1/2}\hat{\boldsymbol{Y}}_k$. Therefore, we first compute the SVD of $\hat{\boldsymbol{A}}_k^{1/2}\hat{\boldsymbol{Y}}_k$ and truncate it to its r principal components, say $\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^T$, where $\boldsymbol{U}_k \in \mathbb{R}^{K\times r}$, $\boldsymbol{V}_k \in \mathbb{R}^{N\times r}$, and $\boldsymbol{S}_k$ is $r\times r$ diagonal with the largest singular values. Then we set

$$\hat{\boldsymbol{L}}_{k+1} = \hat{\boldsymbol{A}}_k^{-1/2}\boldsymbol{U}_k, \qquad \hat{\boldsymbol{R}}_{k+1} = \boldsymbol{S}_k\boldsymbol{V}_k^T. \quad (16)$$

Finally, the value of σ^2 maximizing $Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_k)$ is the one minimizing (13), and is given by

$$\hat{\sigma}_{k+1}^2 = \frac{\frac{1}{N}\left\|\hat{\boldsymbol{A}}_k^{1/2}(\hat{\boldsymbol{Y}}_k - \hat{\boldsymbol{L}}_{k+1}\hat{\boldsymbol{R}}_{k+1})\right\|_F^2 + \hat{\sigma}_k^2\sum_{i=1}^{K}\hat{a}_{i,k}c_i}{\sum_{i=1}^{K}\hat{a}_{i,k}}. \quad (17)$$

4. INITIALIZATION

Since the E-M iteration may in general converge to a local maximum of the likelihood function, its initialization is a critical issue. Next we describe an initialization scheme which appears to be quite effective. It is based on the PCA-based "row-structure fault detection algorithm" from [14], a statistical test to check if a given sensor is faulty. Specifically, let the zero-padded data matrix $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{K\times N}$ be given by

$$\widetilde{\boldsymbol{Y}}_{ij} = y_{ij}, \quad (i,j)\in\Omega, \qquad \widetilde{\boldsymbol{Y}}_{ij} = 0 \quad (i,j)\notin\Omega, \quad (18)$$

and let $\widetilde{\boldsymbol{\mu}}^T = \frac{1}{K}\mathbf{1}_K^T\widetilde{\boldsymbol{Y}}$ be the average of its rows. The $N\times N$ sample covariance matrix is given by $\boldsymbol{C} = \frac{1}{K}(\widetilde{\boldsymbol{Y}} - \mathbf{1}_K\widetilde{\boldsymbol{\mu}}^T)^T(\widetilde{\boldsymbol{Y}} - \mathbf{1}_K\widetilde{\boldsymbol{\mu}}^T)$. Let now $\boldsymbol{\Lambda}_r \in \mathbb{R}^{r\times r}$ be diagonal with the r largest eigenvalues of $\boldsymbol{C}$, and let the columns of $\boldsymbol{U}_r \in \mathbb{R}^{N\times r}$ comprise the corresponding r eigenvectors. According to [14], the test statistic

$$d_i^2 = \|\boldsymbol{\Lambda}_r^{-1/2}\boldsymbol{U}_r^T(\widetilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{\mu}})\|^2, \quad (19)$$

where $\widetilde{\boldsymbol{y}}_i^T$ is the i-th row of $\widetilde{\boldsymbol{Y}}$, is approximately χ_r^2-distributed if the data from sensor i is not abnormal. Using this, the authors of [14] declare sensor i as faulty if d_i^2 exceeds a threshold, which is set to achieve a given probability of false alarm P_{FA}. Since d_i^2 will likely take large values when sensor i is defective yielding corrupted data with high energy, we propose to initialize the probabilities $\hat{a}_{i,0}$ proportionally to these values after normalization. Specifically,

$$\hat{a}_{i,0} = \beta\cdot\frac{d_i^2}{\max_{1\le j\le K}\{d_j^2\}}, \qquad i = 1,\ldots,K, \quad (20)$$

where $0 < \beta \le 1$ is a suitable constant. In practice, $\beta = 0.95$ has been found to give good results. Finally, the initial value $\hat{\boldsymbol{L}}_0\hat{\boldsymbol{R}}_0$ is taken as the best rank-r approximation of $\widetilde{\boldsymbol{Y}}$.

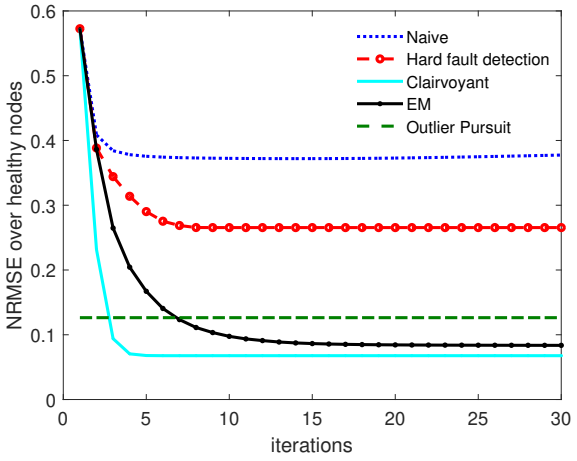

Fig. 1. NRMSE vs. iterations ($r = 10$).

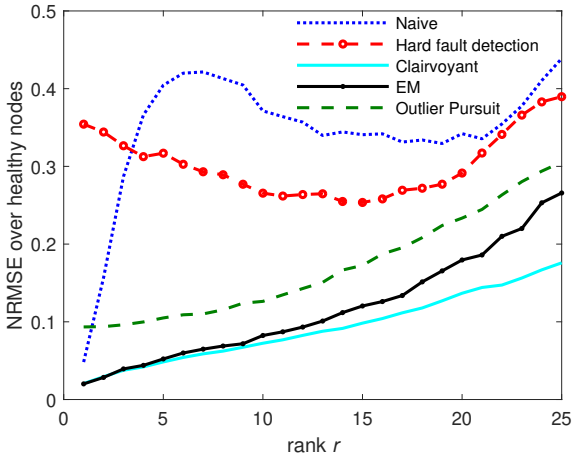

Fig. 2. NRMSE for the different schemes vs. rank r.

5. NUMERICAL RESULTS

Consider a setting with $K = 100$, $N = 70$. Entries of ground truth matrices $\boldsymbol{L} \in \mathbb{R}^{K\times r}$, $\boldsymbol{R} \in \mathbb{R}^{r\times N}$ are independently drawn from a standard Gaussian distribution. The probabilities of sensor malfunction and observing a given data point are 0.1 and 0.65, respectively. The signal-to-noise ratio, defined as $\|\boldsymbol{LR}\|_F^2/(KN\sigma^2)$, is set to 20 dB. The limits of the uniform pdf for faulty sensor data are generated as $A_i = -B_i$ with each B_i independently drawn from a uniform distribution over $[0, 65\sigma]$ to capture a wide range of sensor fault types.

In addition to the proposed E-M estimator, the following schemes were also implemented for comparison:

- A "naive" scheme oblivious to sensor faults, which seeks a minimum of $\|\boldsymbol{y}-\mathcal{P}_\Omega(\boldsymbol{LR})\|_F^2$ by alternatingly optimizing over one of the factors while fixing the other [20], initialized at the best rank-r approximation of the zero-padded matrix $\widetilde{\boldsymbol{Y}}$.

- A "clairvoyant" scheme with knowledge about which sensors are faulty, which applies a similar alternating minimization approach after discarding the corrupted data.

- A scheme based on a preliminary hard fault detection stage declaring sensor i faulty if d_i^2 in (19) exceeds a threshold. Data estimated as corrupted are then discarded before applying the alternating minimization scheme above. As in [14], the threshold is set for $P_{\text{FA}} = 0.025$.

- A scheme based on the "noisy outlier pursuit" method from [23], solving the following convex problem:

$$\min_{\boldsymbol{M},\boldsymbol{S}} \|\boldsymbol{M}\|_* + \lambda\|\boldsymbol{S}^T\|_{1,2} \quad \text{s.to} \quad \|\boldsymbol{y}-\mathcal{P}_\Omega(\boldsymbol{M}+\boldsymbol{S})\| \leq \epsilon, \tag{21}$$

where $\|\cdot\|_*$ is the nuclear norm, and $\|\boldsymbol{B}\|_{1,2}$ is the sum of Euclidean norms of the columns of $\boldsymbol{B}$. In (21), $\boldsymbol{M}$ is an estimate of the low-rank term $\boldsymbol{LR}$, whereas $\boldsymbol{S}$ is an estimate of the corrupted data, with the term $\|\boldsymbol{S}^T\|_{1,2}$ promoting row-sparsity. The tuning parameters were set to $\lambda = 0.9$ and $\epsilon = 0.01$ by trial and error.

For $r = 10$, Fig. 1 shows the evolution of the normalized root mean squared error (NRMSE) over clean data[1], defined as

$$\text{NRMSE} = \frac{\|\boldsymbol{A}(\boldsymbol{LR} - \hat{\boldsymbol{L}}\hat{\boldsymbol{R}})\|_F}{\|\boldsymbol{ALR}\|_F} \tag{22}$$

and averaged over 100 Monte Carlo runs (for the outlier pursuit scheme, $\hat{\boldsymbol{L}}\hat{\boldsymbol{R}}$ in (22) is replaced by $\boldsymbol{M}$). Fig. 2 shows the NRMSE as a function of the ground-truth rank r.

The performance of the clairvoyant estimator constitutes a benchmark. The naive estimator behaves poorly, as could be expected: the effect of corrupted data from faulty sensors cannot be lightly dismissed. In this scenario, the improvement by applying a data cleansing stage as suggested in [14] (hard fault detection) is limited. It was observed that whenever the corrupted data from a faulty sensor has small energy, the fault detection mechanism based on (19) is not effective, but nevertheless the impact on matrix recovery of such low-energy abnormal data remains significant, which explains the poor performance of this scheme. The outlier-pursuit method is seen to provide reasonable performance, although it requires fine tuning of the user-selectable parameters; in addition, its computational load may be a bottleneck as matrix sizes grow. In this sense, we note that the complexity of the proposed E-M scheme is dominated by step (15), which requires the computation of the SVD of a $K \times N$ matrix at each iteration. The E-M estimator gets close to the benchmark, showing the efectiveness of the proposed initialization, with a graceful degradation in performance as the rank r increases.

[1]Corrupted data from faulty sensors is excluded from (22) since it is not possible to recover a whole row of $\boldsymbol{LR}$ from the low-rank property alone.

We repeated the simulations under the same setting, but generating the corrupted data for faulty sensor i as i.i.d. Gaussian with mean $(A_i+B_i)/2$ and variance $\frac{1}{12}(B_i-A_i)^2$, where A_i, B_i were generated in the same way as above. Results were roughly similar to those with uniformly distributed outliers, indicating that E-M is robust to deviations from the proposed non-informative model for corrupted data.

6. CONCLUSIONS

We have presented an ML approach to the problem of matrix completion under noisy data and unreliable sensors. Motivated by the presence of unobserved variables, we developed the E-M iterative algorithm for this problem, together with an effective initialization strategy. E-M outperforms schemes based on hard decisions about the sensor labels, at a fraction of the computational cost of convex optimization approaches, and without requiring parameter tuning.

7. REFERENCES

[1] Yudong Chen and Yuejie Chi, "Harnessing structures in Big Data via guaranteed low-rank matrix estimation," *IEEE Signal Process. Mag.* vol. 35 no. 4, pp. 14-31, Jul. 2018.

[2] M. C. Vuran, Ö. B. Akan, I. F. Akyildiz, "Spatiotemporal correlation: theory and applications for wireless sensor networks", *Computer Networks,* vol. 45 no. 3, pp. 245-259, 2004.

[3] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE J. Sel. Topics Signal Process.,* vol. 10 no. 4, pp. 608-622, Jun. 2016.

[4] Jie Cheng *et al.,* "STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks," *IEEE Trans. Wireless Commun.,* vol. 12 no. 2, pp. 850-861, Feb. 2013.

[5] Jingfei He *et al.,* "Data recovery in wireless sensor networks with joint matrix recovery and sparsity constraints," *IEEE Commun. Lett.,* vol. 19 no. 12, pp. 2230-2233, Dec. 2015.

[6] K. Ni *et al.,* "Sensor network data fault types," *ACM Trans. Sensor Netw.,* vol. 5 no. 3, May 2009.

[7] A. Mahapatro and P. M. Khilar, "Fault diagnosis in wireless sensor networks: a survey," *IEEE Commun. Surv. Tuts.,* vol. 15 no. 4, pp. 2000-2026, 2013.

[8] X. Du and H.-H. Chen, "Security in wireless sensor networks," *IEEE Trans. Wireless Commun.,* vol. 15 no. 4, pp. 60-66, Aug. 2008.

[9] A. Vempaty, L. Tong and P. Varshney, "Distributed inference with Byzantine data: state of the art review on data falsification attacks," *IEEE Signal Process. Mag.,* vol. 30 no. 5, pp. 65-75, Sep. 2013.

[10] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge, U.K.: Cambridge Univ. Press, 1985.

[11] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions.* New York, NY, USA: Wiley, 1997.

[12] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. 20th Int. Conf. Machine Learning,* pp. 720-727, 2003.

[13] P. M. Q. Aguiar and J. M. F. Moura, "Factorization as a rank 1 problem," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition,* vol. 1, pp. 178-184, 1999.

[14] K. Xie *et al.,* "Recover corrupted data in sensor networks: a matrix completion solution," *IEEE Trans. Mobile Comput.,* vol. 16 no. 5, pp. 1434-1448, May 2017.

[15] L. L. Scharf, *Statistical signal processing: detection, estimation, and time series analysis.* Reading, MA, USA: Addison-Wesley; 1990.

[16] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions.* New York, NY, USA: Wiley; 1997.

[17] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory,* vol. 56 no. 5, pp. 2053-2080, May 2010.

[18] B. Recht, M. Fazel and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.,* vol. 52 no.3, pp. 471-501, 2010.

[19] S. Burer and R. D. C. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Programming,* vol. 103 no. 3, pp. 427-444, Jul. 2005.

[20] M. Hardt, "Understanding alternating minimization for matrix completion," in *Proc. IEEE 55th Annu. Symp. Found. Comput. Sci.,* pp. 651-660, 2014.

[21] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Proc. 28th Conf. Learning Theory,* pp. 1007-1034, 2015.

[22] E. J. Candès *et al.,* "Robust Principal Component Analysis?," *J. of the ACM,* vol. 58 no. 3, art. 11 , May 2011.

[23] H. Xu, C. Caramanis and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Trans. Inf. Theory,* vol. 58 no. 5, pp. 3047-3064, May 2012.

[24] Yudong Chen *et al.,* "Low-rank matrix recovery from errors and erasures," *IEEE Trans. Inf. Theory,* vol. 59 no. 7, pp. 4324-4337, Jul. 2013.