Sensor Data Reconstruction in Industrial Environments with Cellular Connectivity

Charalampos Kalalas and Jesus Alonso-Zarate
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Barcelona, Spain
Emails: {ckalalas, jesus.alonso}@cttc.es

***Abstract*—The reliable acquisition of monitoring information is critical for several industrial use cases relying on wireless sensor network deployments. However, missing sensor measurements are typical in industrial systems empowered by cellular connectivity due to the stochastic nature of the wireless channel. In this paper, we propose a sensor data reconstruction scheme that exploits the hidden data dynamics to accurately estimate the missing measurements. Based on an analytical framework for the network model and a closed-form expression for the outage probability, the impact on the reconstruction error performance is thoroughly explored. Considering a dataset with high spatiotemporal correlation in the sensor observations, our proposed scheme is shown to outperform two baseline data recovery methods in terms of reconstruction error for various network configurations. In addition, despite the presence of imperfect cellular connectivity, our proposed scheme exhibits high event-detection accuracy.**

***Index Terms*—time series, missing data, expectation maximization, parameter learning, industrial IoT, uplink communication**

I. Introduction

The ongoing modernization of the aging industrial systems aims at the transformation of the entire production chain, composed of processes, machines, workers, and production lines, into a fully integrated, automated and interconnected paradigm to improve productivity and operational efficiency. This evolution resides at the core of the upcoming Industry 4.0 era and heavily relies on the digitalization and ubiquitous connectivity provided by 5G mobile systems and Internet of Things (IoT) technologies [1]–[3]. The integration of 5G ultra-reliable low-latency communication (URLLC) as well as the massive deployment of intelligent wireless sensors are expected to: *i*) unlock unprecedented industrial use cases, e.g., non-stationary machines, scalable line configuration, etc.; *ii*) allow the expansion of digital operations, e.g., for remote plant condition monitoring; and *iii*) contribute to the extension of machine lifetime, e.g., via predictive maintenance and timely anomaly detection.

Instrumental to the ongoing industrial transformation is the availability and accessibility of well-structured datasets related with the monitoring and condition maintenance of industrial equipment [4]. However, in industrial IoT systems empowered by wireless connectivity, the unreliable nature of the wireless medium constitutes one of the fundamental challenges introducing uncertainties in various state-monitoring operations. The complex fading conditions in factory plants, usually rich in metallic surfaces and physical obstructions, result in high network dynamics and harsh radio propagation environments which significantly affect the accuracy of the transmitted sensor measurements and often result in transmission failures. In addition, industrial datasets often exhibit high spatiotemporal correlation with context dependencies that render the data characterization and modeling tasks non-trivial.

During the recent years, time series analysis has attracted considerable research attention for the accurate detection, identification and diagnosis of abnormal behaviours, e.g., faulty conditions, disturbances, etc., in industrial environments [5]–[12]. In the majority of existing works, however, the stochastic nature of the wireless channel and its impact on the sensor measurements is either neglected [5]–[9] or limitedly considered [10]–[12]. In fact, it is a common approach in the literature to consider simplified models for the wireless channel reliability; either with the use of a generic expression for the probability of failure, e.g., as in [10], [11], or with the aid of indicator functions for the unsuccessful transmissions, e.g., as in [12], and without explicit network topology considerations. Instead, a more rigorous analysis of the communication impairments is necessary for a thorough assessment of the impact of various network parameters on the quality of the sensor measurements. The authors in [13] employ a Kalman filter based state estimation model for industrial automation applications and propose an optimized small-packet transmission scheme to improve the network-wide revenue. Even though erroneous sensor transmissions are incorporated in the system model, missing data reconstruction is not explicitly addressed and the focus is solely on the optimization of the small-packet transmission scheme. In addition, to the best of the authors' knowledge, there is no prior work studying the recovery of missing sensor measurements transmitted over a wireless channel under the presence of spatiotemporal dynamics in the sensors' observations.

Contribution: Motivated by these literature gaps, our contribution in this paper is twofold:

- We propose a tractable analytical framework to incorporate the imperfect cellular connectivity in the sensor data reconstruction process. Our framework relies on stochastic geometry modeling principles, able to deal with networks of random topologies. A closed-form expression for the outage probability allows for an accurate representation of an incomplete sensor data sequence with missing measurements due to the unreliable nature of the wireless channel.
- We develop a data reconstruction scheme that takes into

account the hidden dynamics, i.e., spatiotemporal context dependencies, in the dataset to estimate the missing sensor measurements. The performance assessment in terms of reconstruction error reveals significant reconstruction accuracy gains against baseline methods and offers useful insights for the design of industrial IoT systems empowered by cellular connectivity.

Organization: The rest of the paper is organized as follows. The system model, assumptions and the methodology for the connectivity analysis are presented in Section II. Section III describes in detail the proposed missing data reconstruction scheme along with all the involved steps of an iterative expectation-maximization (EM) algorithm employed for the estimation of the missing sensor measurements. Model validation and performance comparison of our proposed scheme against baseline methods for missing data recovery are presented in Section IV. Finally, Section V concludes the paper.

II. System Model

A. Network Model

We consider an industrial process monitoring scenario composed of a data fusion center and a number of deployed sensors able to monitor the operational state and condition of the various components. The sensors are equipped with radio interfaces and transmit their state measurements to the fusion center which is co-located with their associated base station[1] (BS). A stochastic geometry framework is considered for the positions of the BSs and the sensors. In particular, we assume that the BSs are distributed on the plane according to a homogeneous Poisson point process (PPP) Ψ_b of intensity λ_b, while the sensors are spatially distributed according to a homogeneous PPP $\Psi_u = \{u_k; i = 1, 2, 3, \dots\}$ with intensity λ_u. We further assume saturated uplink traffic conditions where all sensors always have data to transmit while λ_u is considered high enough such that each BS will serve at least one sensor per channel.

B. Channel Model and Power Control

Regarding the channel model, a generic power-law path-loss model is considered where the signal power decays with the propagation distance r at the rate $r^{-\alpha}$, where α denotes the path-loss exponent. In addition to the path-loss attenuation, the channel power gains are assumed to be independent of each other and of the spatial locations, and identically distributed (i.i.d). We assume Rayleigh fast fading where the channel power gains, denoted by h, are exponentially distributed with unit mean. Let p_k denote the transmit power of sensor k. All sensors are considered to have equal maximum transmit power denoted by $p_{\max}$ and employ a truncated channel inversion power control policy to compensate for the path loss and keep the average received signal power at the BS equal to a threshold ρ which is greater than the receiver sensitivity $\rho_{\min}$.

[1]A closest BS association scheme is adopted in this paper.

C. Outage Probability

If the transmit power required for the path-loss inversion is higher than $p_{\max}$, the sensor is considered to be in outage due to insufficient transmit power and does not transmit its measurement. Based on the transmit power analysis conducted in [14], the probability that a sensor experiences outage due to insufficient transmit power for $\alpha = 4$ is given by

$$P_{o_1} = e^{-\pi\lambda_b\sqrt{\frac{p_{\max}}{\rho}}}. \tag{1}$$

From Eq. (1), it can be observed that P_{o_1} depends on the BS intensity λ_b, the maximum sensor transmit power $p_{\max}$, and the threshold ρ which constitutes a network design parameter.

A measurement from an active sensor (i.e., a sensor not in power outage) is considered to be successfully decoded at the serving BS if and only if the signal-to-interference-plus-noise ratio (SINR) of the useful signal is greater than a certain threshold γ_{th}. Based on the stationarity of the PPP, the SINR experienced at the origin BS for a target active sensor can be expressed as

$$\text{SINR} = \frac{\rho h_0}{\sigma^2 + \sum_{u_k \in \tilde{\Psi}_u} p_k h_k \left\|u_k\right\|^{-\alpha}}, \tag{2}$$

where the numerator corresponds to the useful signal power, σ^2 is the noise power while the second term in the denominator denotes the aggregate interference from other uplink sensor transmissions on the same channel[2]. Then, according to the SINR analysis conducted in [14] and for $\alpha = 4$, the SINR outage probability $P_{o_2} = \mathbb{P}(\text{SINR} < \gamma_{\text{th}})$ for an active sensor can be calculated as

$$P_{o_2} = 1 - \exp\left\{-\frac{\gamma_{\text{th}}\sigma^2}{\rho} - \frac{\sqrt{\gamma_{\text{th}}}\gamma\left(2, \pi\lambda_b\sqrt{\frac{p_{\max}}{\rho}}\right)}{\left(1 - e^{-\pi\lambda_b\sqrt{\frac{p_{\max}}{\rho}}}\right)}\arctan(\sqrt{\gamma_{\text{th}}})\right\}, \tag{3}$$

where $\gamma(a, b) = \int_0^b t^{a-1}e^{-t}dt$ denotes the lower incomplete gamma function. The total outage probability can then be expressed as

$$P_o = P_{o_1} + (1 - P_{o_1})\, P_{o_2}, \tag{4}$$

where P_{o_1} and P_{o_2} can be calculated using Eqs. (1) and (3) respectively.

III. Sensor Data Reconstruction

At the fusion center located at each serving BS, the received measurements can be represented by a partially observable time sequence $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$, where each vector $\mathbf{y}_t$ contains the received measurements at time-step t from the deployed sensors. In particular, by incorporating the outage probability P_o (given by Eq. (4)), the entries $y_{t,k}$ of $\mathbf{y}_t$ can be expressed as

$$y_{t,k} = (1 - P_o)x_{t,k} + v_{t,k}, \tag{5}$$

[2]For analytical tractability, the locations of the interfering sensors are considered to follow a PPP and their transmit powers p_k are independent.

where $x_{t,k}$ corresponds to the transmitted measurement from sensor k at time-step t while $\upsilon_{t,k}$ is a zero-mean white Gaussian noise with covariance matrix Ω. It is noted that Eq. (5) takes into account the stochastic nature of the wireless channel which may result in a received vector $\mathbf{y}_t$ with intermittent measurements.

We consider a time sequence of latent variables (i.e., hidden states) $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T]$ to model the dynamics and the hidden patterns of the received measurements. We also introduce an indicator matrix, Φ, for the missing measurements; i.e., $\phi_{t,k} = 0$ whenever the k-th sensor measurement in $\mathbf{y}_t$ is missing at time t; otherwise, $\phi_{t,k} = 1$. Let us also denote the observed part of $\mathbf{Y}$ as $\mathbf{Y}_r$ and the missing part as $\mathbf{Y}_m$. Following the rationale of linear dynamical systems [15], our model for the received measurements at the fusion center can be described by the following two equations:

$$\mathbf{z}_{t+1} = A\mathbf{z}_t + \mathbf{w}_t, \tag{6}$$

$$\mathbf{y}_t = C\mathbf{z}_t + \mathbf{v}_t. \tag{7}$$

To capture temporal correlation, we assume that the latent variables at each time tick depend linearly on the previous values via the linear state transition matrix A. At each time tick, the received vector $\mathbf{y}_t$, including both observed and missing sensor measurements, is assumed to be a linear function of the latent variables $\mathbf{z}_t$ via the linear projection matrix C. This mapping implicitly captures the spatial correlation among the different sensor measurements [16]. Both hidden state evolution and received measurement processes are corrupted by zero-mean white Gaussian noise, $\mathbf{w}_t$ and $\mathbf{v}_t$, with covariance matrices, Q and R, respectively. Further, $\mathbf{w}_t$ and $\mathbf{v}_t$ are assumed to be independent. The initial state $\mathbf{z}_0$ of the latent variables is also a Gaussian random variable with mean $\boldsymbol{\pi_1}$ and covariance V_1. Therefore, the parameter vector of our model is $\theta = (A, C, Q, R, \boldsymbol{\pi}_1, V_1)$.

Based on Eqs. (6) and (7), we can express the conditional probabilities for the hidden state and the received sequence, respectively, as follows:

$$P\left(\mathbf{z}_t|\mathbf{z}_{t-1}\right) = \exp\left\{-\tfrac{1}{2}D\left(\mathbf{z}_t, A\mathbf{z}_{t-1}, Q\right)\right\}(2\pi)^{-\kappa_1/2}|Q|^{-1/2}, \tag{8}$$

$$P\left(\mathbf{y}_t|\mathbf{z}_t\right) = \exp\left\{-\tfrac{1}{2}D\left(\mathbf{y}_t, C\mathbf{z}_t, R\right)\right\}(2\pi)^{-\kappa_2/2}|R|^{-1/2}, \tag{9}$$

where $D\left(\boldsymbol{\omega}_t, \boldsymbol{\mu}_t, \Xi\right) = \left(\boldsymbol{\omega}_t - \boldsymbol{\mu}_t\right)' \Xi^{-1}\left(\omega_t - \boldsymbol{\mu}_t\right)$ denotes the square of the Mahalanobis distance of a vector $\boldsymbol{\omega}_t$ with mean vector $\boldsymbol{\mu}_t$ and covariance matrix Ξ.

Based on the Markov property implicit in the model, the factored representation of the joint probability distribution of $\mathbf{Z}$ and $\mathbf{Y}$ is given by

$$P\left(\mathbf{Z}, \mathbf{Y}|\theta\right) = P\left(\mathbf{z}_1\right)\prod_{t=2}^{T} P\left(\mathbf{z}_t|\mathbf{z}_{t-1}\right)\prod_{t=1}^{T} P\left(\mathbf{y}_t|\mathbf{z}_t\right), \tag{10}$$

and the joint log-likelihood can be written as

$$\begin{aligned}\log P\left(\mathbf{Z}, \mathbf{Y}|\theta\right) = &-\frac{1}{2}\left(D\left(\mathbf{z}_1, \boldsymbol{\pi}_1, V_1\right) - \log|V_1| - T(\kappa_1 + \kappa_2)\log 2\pi\right) \\ &- \sum_{t=2}^{T}\left(\frac{1}{2}D\left(\mathbf{z}_t, A\mathbf{z}_{t-1}, Q\right)\right) - \frac{T-1}{2}\log|Q| \\ &- \sum_{t=1}^{T}\left(\frac{1}{2}D\left(\mathbf{y}_t, C\mathbf{z}_t, R\right)\right) - \frac{T}{2}\log|R|.\end{aligned} \tag{11}$$

Given that the received sequence $\mathbf{Y}$ is characterized by intermittent measurements due to the imperfect cellular connectivity, our goal is to maximize the conditional expectation of the received data log-likelihood, i.e.,

$$L\left(\theta\right) = E_{\mathbf{Y}_m, \mathbf{Z}|\mathbf{Y}_r, \Phi}[\log P(\mathbf{Z}, \mathbf{Y}|\theta)]. \tag{12}$$

To that aim, we apply an iterative EM algorithm following a coordinate descent procedure [17]. We provide the details in the following subsection.

A. The EM algorithm

The EM algorithm provides an iterative method for finding the maximum likelihood estimates of θ based on the observed measurements, $\mathbf{Y}_r$, by successively maximizing Eq. (12). The E-step of EM algorithm requires computing $L\left(\theta\right)$ in Eq. (12). Based on Eq. (11), this computation amounts to deriving the following three expectations:

$$\hat{\mathbf{z}}_t \equiv E\left[\mathbf{z}_t|\mathbf{Y}\right], \tag{13}$$

$$P_t \equiv E\left[\mathbf{z}_t\mathbf{z}_t'|\mathbf{Y}\right], \tag{14}$$

$$P_{t,t-1} \equiv E\left[\mathbf{z}_t\mathbf{z}_{t-1}'|\mathbf{Y}\right]. \tag{15}$$

Let $\mathbf{z}_t^\tau$ and V_t^τ denote $E\left(\mathbf{z}_t|Y_1^\tau\right)$ and $\mathrm{Var}\left(\mathbf{z}_t|Y_1^\tau\right)$, respectively, for the subsequence of received measurements until time τ. Note that $\mathbf{z}_0^1 = \boldsymbol{\pi}_1$ and $V_0^1 = V_1$. Let also θ be an initialization of the parameter vector. The conditional expectations in Eqs. (13)–(15) can be expressed as

$$\hat{\mathbf{z}}_t = \mathbf{z}_t^T, \tag{16}$$

$$P_t = V_t^T + \mathbf{z}_t^T\mathbf{z}_t^{T\prime}, \tag{17}$$

$$P_{t,t-1} = V_{t,t-1}^T + \mathbf{z}_t^T\mathbf{z}_{t-1}^{T\ \prime}. \tag{18}$$

and their computation can be decomposed into the following sets of forward and backward recursion:

i) Forward recursion:

$$\mathbf{z}_t^{t-1} = A\mathbf{z}_{t-1}^{t-1}, \tag{19}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A' + Q, \tag{20}$$

$$K_t = V_t^{t-1}C'\left(CV_t^{t-1}C' + R\right)^{-1}, \tag{21}$$

$$\mathbf{z}_t^t = \mathbf{z}_t^{t-1} + K_t\left(\mathbf{y}_t - C\mathbf{z}_t^{t-1}\right), \tag{22}$$

$$V_t^t = V_t^{t-1} - K_tCV_t^{t-1}. \tag{23}$$

ii) Backward recursion:

$$J_{t-1} = V_{t-1}^{t-1} A' \left(V_t^{t-1}\right)^{-1}, \tag{24}$$

$$\mathbf{z}_{t-1}^{T} = \mathbf{z}_{t-1}^{t-1} + J_{t-1}\left(\mathbf{z}_t^{T} - A\mathbf{z}_{t-1}^{t-1}\right), \tag{25}$$

$$V_{t-1}^{T} = V_{t-1}^{t-1} + J_{t-1}\left(V_t^{T} - V_t^{t-1}\right) J_{t-1}', \tag{26}$$

$$V_{t-1,t-2}^{T} = V_{t-1}^{t-1} J_{t-2}' + J_{t-1}\left(V_{t,t-1}^{T} - AV_{t-1}^{t-1}\right) J_{t-2}', \tag{27}$$

where Eq. (27) is initialized as $V_{T,T-1}^{T} = (I - K_T C) AV_{T-1}^{T-1}$.

After calculating the conditional expectations of the latent variables (E-step), the M-step re-estimates the parameter vector θ to be used in the E-step. To estimate $\theta = (A, C, Q, R, \boldsymbol{\pi}_1, V_1)$, we take the respective partial derivative of Eq. (12), set to zero, and solve for the value of each respective parameter. In particular, the updated parameters are computed as follows:

i) Projection matrix:

$$\frac{\partial L}{\partial C} = -\sum_{t=1}^{T} R^{-1}\mathbf{y}_t\hat{\mathbf{z}}_t' + \sum_{t=1}^{T} R^{-1} C P_t = 0,$$

$$C^{\text{new}} = \left(\sum_{t=1}^{T}\mathbf{y}_t\hat{\mathbf{z}}_t'\right)\left(\sum_{t=1}^{T} P_t\right)^{-1}. \tag{28}$$

ii) Measurement noise covariance:

$$\frac{\partial L}{\partial R^{-1}} = \frac{T}{2}R - \sum_{t=1}^{T}\left(\frac{1}{2}\mathbf{y}_t\mathbf{y}_t' - C\hat{\mathbf{z}}_t\mathbf{y}_t' + \frac{1}{2}CP_tC'\right) = 0,$$

$$R^{\text{new}} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{y}_t\mathbf{y}_t' - C^{\text{new}}\hat{\mathbf{z}}_t\mathbf{y}_t'\right). \tag{29}$$

iii) State transition matrix:

$$\frac{\partial L}{\partial A} = -\sum_{t=2}^{T} Q^{-1}P_{t,t-1} + \sum_{t=2}^{T} Q^{-1}AP_{t-1} = 0,$$

$$A^{\text{new}} = \left(\sum_{t=2}^{T} P_{t,t-1}\right)\left(\sum_{t=2}^{T} P_{t-1}\right)^{-1}. \tag{30}$$

iv) State noise covariance:

$$\frac{\partial L}{\partial Q^{-1}} = \frac{T-1}{2}Q - \frac{1}{2}\left(\sum_{t=2}^{T} P_t - A^{\text{new}}\sum_{t=2}^{T} P_{t-1,t}\right) = 0,$$

$$Q^{\text{new}} = \frac{1}{T-1}\left(\sum_{t=2}^{T} P_t - A^{\text{new}}\sum_{t=2}^{T} P_{t-1,t}\right). \tag{31}$$

v) Initial state mean:

$$\frac{\partial L}{\partial \boldsymbol{\pi}_1} = V_1^{-1}\left(\hat{\mathbf{z}}_1 - \boldsymbol{\pi}_1\right) = 0,$$

$$\boldsymbol{\pi}_1^{\text{new}} = \hat{\mathbf{z}}_1. \tag{32}$$

vi) Initial state covariance:

$$\frac{\partial L}{\partial V_1^{-1}} = \frac{1}{2}V_1 - \frac{1}{2}\left(P_1 - \hat{\mathbf{z}}_1\boldsymbol{\pi}_1' - \boldsymbol{\pi}_1\hat{\mathbf{z}}_1' + \boldsymbol{\pi}_1\boldsymbol{\pi}_1'\right) = 0,$$

$$V_1^{\text{new}} = P_1 - \hat{\mathbf{z}}_1\hat{\mathbf{z}}_1'. \tag{33}$$

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Path-loss exponent α	4
Maximum transmit power $p_{\max}$	1W
BS density λ_b	10BSs/km^2
Threshold ρ	-70dBm
Noise power σ^2	-90dBm
SINR threshold γ_{th}	0dBm

Finally, using the Markov property, the missing sensor measurements $\mathbf{Y}_m$ can be computed from the estimation of the latent variables as

$$E[\mathbf{Y}_m|\mathbf{Y}_r, \mathbf{Z}; \theta] = C^{\text{new}}\, E[Z|_{(t,k)},\, \phi_{t,k} = 0]. \tag{34}$$

The Eqs. (13)–(27) (E-step) and Eqs. (28)–(33) (M-step) complete one iteration of the EM algorithm; these equations are alternated repeatedly until the difference $L\left(\theta^{\text{new}}\right) - L\left(\theta^{\text{old}}\right)$ changes by an arbitrary small amount ϵ.

IV. Results and Discussion

In this section, we aim to validate our proposed analytical framework against simulation results and provide a performance comparison in terms of reconstruction error of our proposed scheme with respect to two baseline methods. For the simulation setup, we consider a PPP for the BSs with intensity λ_b and sensors are randomly spread over the area until all BSs are serving at least one sensor, i.e., to achieve traffic saturation conditions. Table I summarizes the parameter values used in our simulations, unless otherwise stated.

The considered dataset consists of IEC-61850 network traffic traces generated by intelligent electronic devices based on open-source libraries for protocol implementation[3] and a discrete-event network simulator [18]. The traffic streams correspond to both intra- and inter-substation communication in power systems and represent both normal and disturbance scenarios. The dataset exhibits high levels of spatial and temporal correlation across the device measurements to model the real behaviour of substation automation systems, e.g., local cascade failures result in inter-dependent transmissions of measurements from neighboring devices.

For performance comparison, we consider the following two baseline data reconstruction schemes: *i*) a linear interpolation method which exploits temporal continuity to compute the missing value estimates using observed sensor measurements in neighboring time ticks; and *ii*) a method based on Singular Value Decomposition (SVD) which transforms the weighted low-rank approximation problem to a maximum-likelihood problem with missing values and approximates them by computing the SVD iteratively [19]. In our proposed scheme, we initialize our estimated received time sequence $\hat{\mathbf{Y}}$ with $\mathbf{Y}_r$ while the missing sensor measurements are initially reconstructed by means of linear interpolation and then iteratively imputed as in Eq. (34). The process continues by updating the

[3]Available online at: https://github.com/mz-automation/libiec61850

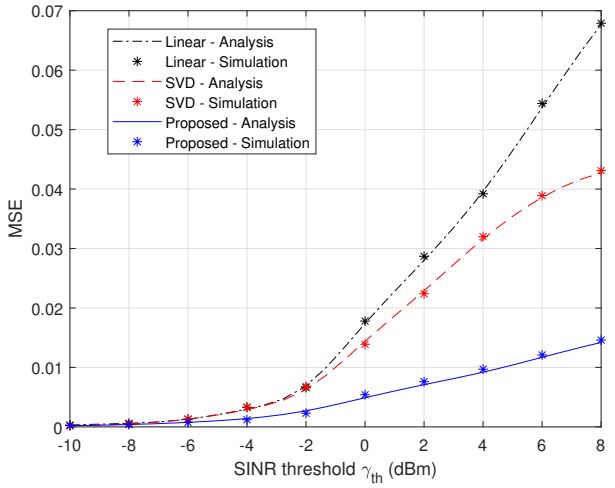

Fig. 1. Performance comparison of the different missing data reconstruction schemes in terms of reconstruction error for varying SINR threshold γ_{th}.

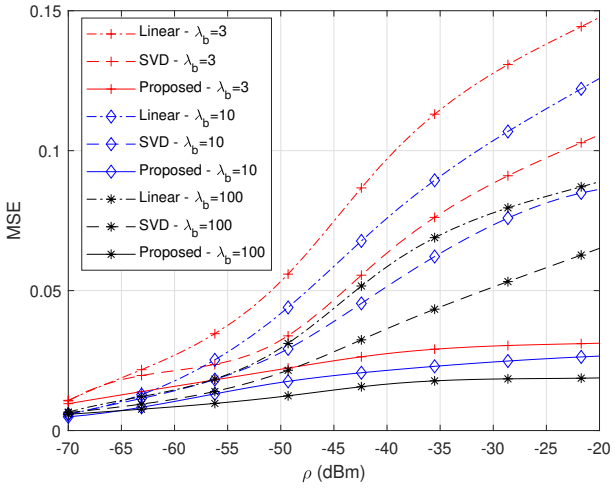

Fig. 2. Performance comparison of the different missing data reconstruction schemes in terms of reconstruction error for varying ρ and λ_b.

expectations of the latent variables based on the newly imputed values of the missing measurements until convergence. We further assume that the noise covariances in θ constitute diagonal matrices.

The effectiveness of reconstruction is evaluated in terms of the mean squared error (MSE), defined as the average of the squared differences between the real and reconstructed missing measurements, i.e.,

$$\text{MSE}(\mathbf{Y}, \Phi, \hat{\mathbf{Y}}) = \frac{1}{\sum_{t,k}(1-\phi_{t,k})} \sum_{t,k} (1-\phi_{t,k}) \left(Y_{t,k} - \hat{Y}_{t,k}\right)^2. \tag{35}$$

To reduce random effects, we repeat each simulation 10000 times and we report the average of the MSE.

Fig. 1 illustrates the MSE performance with respect to the SINR threshold γ_{th} for different reconstruction schemes. It can be observed that the MSE increases with increasing SINR threshold γ_{th}, as the aggregate interference leads to significant signal degradation and results in a high number of sensor links in SINR outage. In addition, the increased outage probability in the high γ_{th} regime, impacts the achieved MSE and may lead to lower performance of the data reconstruction process. However, in stark contrast with the linear interpolation and SVD-based methods, our proposed scheme keeps the MSE in relatively lower levels which is especially important for high values of γ_{th}. We can also observe a tight match between the analytical curves and the simulation results which validates the accuracy of the outage probability expression in Eq. (4).

Fig. 2 depicts the evolution of the MSE with respect to various network configurations for the three missing data reconstruction schemes. It can be observed that MSE increases with increasing ρ, as a higher received power threshold at the BS leads to an increased number of lost sensor measurements. This can be intuitively explained as follows. When ρ increases, the sensors are required to transmit at higher power levels; this, in turn, leads to an increasing number of sensors failing to establish an uplink connection with their serving BS due to

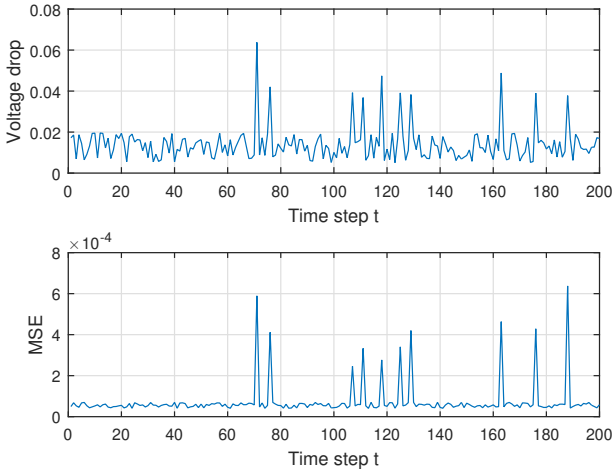

Fig. 3. Event detection accuracy of our proposed scheme. The MSE peaks (bottom figure) nearly coincide with the time ticks of the event injections in the dataset (top figure).

insufficient transmit power to invert their path loss, given that p_{max} is considered constant (refer also to Eq. (1)). It is worth noting that the achieved MSE levels of our proposed scheme remain significantly lower with respect to the linear interpolation and SVD-based methods. In addition, data reconstruction performance improves with increasing BS intensity λ_b. When the BSs are dense enough, the distance between a generic sensor and its corresponding serving BS decreases and the required transmit power for path-loss inversion is lower. It is worth noting that the data reconstruction gains of our scheme can be capitalized even for low values of λ_b, demonstrating the robustness of our approach even for a sparse deployment of BSs.

Finally, Fig. 3 illustrates the event-detection performance of our proposed scheme. In particular, as shown in the top

plot of Fig. 3, we have artificially "injected" various events in the dataset, corresponding to abnormal conditions, i.e., voltage drop beyond a maximum allowable level of 3%. The superposition of multiple events in the dataset generates different patterns which translate to different model parameters θ and latent variables $\mathbf{Z}$. Based on its property to compute the posterior distribution of the latent variables in each time tick, our proposed scheme is able to accurately capture the dataset dynamics, i.e., events, by tracking the peak values of the MSE shown in the bottom plot of Fig. 3. It is worth noting that the MSE peaks nearly coincide with the time ticks of the events in the top plot of Fig. 3, despite the presence of imperfect connectivity.

V. Conclusions

Industrial systems empowered by cellular connectivity heavily rely on the reliable acquisition of sensor measurements. However, due to the wireless channel impairments, the sensor measurements received at the fusion center are often highly intermittent. In this paper, a reconstruction method for the estimation of missing sensor measurements has been proposed, considering a stochastic geometry framework for the network topology in industrial environments. Based on a closed-form expression for the outage probability, the impact on the missing values' estimation accuracy is explored under different network configurations with the aid of a dataset which consists of spatiotemporal sensor measurements. The simulation results demonstrate the superior performance of our proposed scheme against two baseline data recovery approaches, while the achieved event-detection performance is shown to be highly accurate despite the incomplete received data.

Future work will aim at extending the proposed framework to capture non-linear and multi-scale spatiotemporal behavior of the sensor measurements by resorting to the utilization of switching linear Gaussian dynamical systems. In addition, we plan to further investigate the applicability of our proposed scheme for other data mining tasks, e.g., for data compression when the storing capabilities of the fusion center impose limitations on the amount of stored sensor data.

Acknowledgment

This work has been partially supported by the CHIST-ERA FIREMAN project funded by the Spanish national foundation (PCI2019-103780), by the Spanish MINECO project SPOT5G (TEC2017-87456-P), and by the Generalitat de Catalunya under Grant 2017 SGR 891.

References

[1] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The Future of Industrial Communication: Automation Networks in the Era of the Internet of Things and Industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, 2017.

[2] D. Gutierrez-Rojas, M. Ullah, I. T. Christou, G. Almeida, P. H. J. Nardelli, D. Carrillo, J. M. Sant'Ana, H. Alves, M. Dzaferagic, A. Chiumento, and C. Kalalas, "Three-layer Approach to Detect Anomalies in Industrial Environments based on Machine Learning," 2020.

[3] J. M. d. S. Sant'Ana, A. Hoeller, R. D. Souza, S. Montejo-Sánchez, H. Alves, and M. d. Noronha-Neto, "Hybrid Coded Replication in LoRa Networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5577–5585, 2020.

[4] X. Jiang, Z. Pang, M. Luvisotto, F. Pan, R. Candell, and C. Fischione, "Using a Large Data Set to Improve Industrial Wireless Communications: Latency, Reliability, and Security," *IEEE Industrial Electronics Magazine*, vol. 13, no. 1, pp. 6–12, 2019.

[5] D. Zurita, M. Delgado, J. A. Carino, J. A. Ortega, and G. Clerc, "Industrial Time Series Modelling by Means of the Neo-Fuzzy Neuron," *IEEE Access*, vol. 4, pp. 6151–6160, 2016.

[6] D. Zurita, M. Delgado, J. A. Carino, and J. A. Ortega, "Multimodal Forecasting Methodology Applied to Industrial Process Monitoring," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 494–503, 2018.

[7] S. Askari, N. Montazerin, and M. H. F. Zarandi, "High-Frequency Modeling of Natural Gas Networks From Low-Frequency Nodal Meter Readings Using Time-Series Disaggregation," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 136–147, 2016.

[8] T. Chen, X. Liu, B. Xia, W. Wang, and Y. Lai, "Unsupervised Anomaly Detection of Industrial Robots Using Sliding-Window Convolutional Variational Autoencoder," *IEEE Access*, vol. 8, pp. 47072–47081, 2020.

[9] S. Yin, J. J. Rodriguez-Andina, and Y. Jiang, "Real-Time Monitoring and Control of Industrial Cyberphysical Systems: With Integrated Plant-Wide Monitoring and Control Framework," *IEEE Industrial Electronics Magazine*, vol. 13, no. 4, pp. 38–47, 2019.

[10] P. Hu and J. Zhang, "5G-Enabled Fault Detection and Diagnostics: How Do We Achieve Efficiency?," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3267–3281, 2020.

[11] D. Gunatilaka, M. Sha, and C. Lu, "Impacts of channel selection on industrial wireless sensor-actuator networks," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, 2017.

[12] L. Liu and W. Yu, "A D2D-Based Protocol for Ultra-Reliable Wireless Communications for Industrial Automation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5045–5058, 2018.

[13] L. Lyu, C. Chen, S. Zhu, and X. Guan, "5G Enabled Codesign of Energy-Efficient Transmission and Estimation for Industrial IoT Systems," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 6, pp. 2690–2704, 2018.

[14] H. ElSawy and E. Hossain, "On Stochastic Geometry Modeling of Cellular Uplink Transmission With Truncated Channel Inversion Power Control," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4454–4469, 2014.

[15] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2005.

[16] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," tech. rep., USA, 1995.

[17] S. Roweis and Z. Ghahramani, "A Unifying Review of Linear Gaussian Models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[18] C. Kalalas, J. Alonso-Zarate, and G. Bag, "On the Transmission Mode Selection for Substation Automation Traffic in Cellular Networks," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 1–7, October 2017.

[19] N. Srebro and T. Jaakkola, "Weighted Low-Rank Approximations," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, p. 720–727, AAAI Press, 2003.