IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 24, 2018, accepted July 27, 2018, date of publication August 7, 2018, date of current version September 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2864126

Data Reconstruction in Wireless Sensor Networks From Incomplete and Erroneous Observations

ZHENGYU CHEN[1], LEI CHEN[2,3], GUOBING HU[1], WENCAI YE[3], JIN ZHANG[1], AND GENG YANG[3]

[1]School of Electronic and Information Engineering, Jinling Institute of Technology, Nanjing 211169, China
[2]Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210023, China
[3]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Corresponding author: Zhengyu Chen (zych@jit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872190 and Grant 61572263, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20161516, Grant BK20130096, and Grant BK20161104, in part by the China Postdoctoral Science Foundation under Grant 2015M581794, and in part by the Postdoctoral Science Foundation of Jiangsu Province under Grant 1501023C.

ABSTRACT Many basic scientific works use wireless sensor networks (WSNs) to collect environmental data and use the observations for scientific research. The completeness and accuracy of the collected environmental observations determine the reliability of the research results. However, due to the inherent characteristics of WSNs, data loss, and data error usually occur during the process of data collection. Therefore, it is necessary to design an effective method to reconstruct the environmental data from the incomplete and erroneous observations. In this paper, we propose a novel data reconstruction scheme via temporal stability guided matrix completion. First, based on the low-rank feature of sensory environmental data, we formulate the data reconstruction problem as a matrix completion with structural noise. We also introduce a constraint about short-term stability to the matrix completion problem for further reducing the reconstruction error. We then, design an algorithm based on the block coordinate descent method and the operator splitting technique to solve the problem. Finally, simulation results on real sensory data sets show that the proposed approach not only significantly outperforms existing solutions in terms of reconstruction accuracy but also can recognize the sensor nodes with erroneous sensory data.

INDEX TERMS Wireless sensor networks, data collection, matrix completion, data reconstruction.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) have been widely used in many applications, including environmental monitoring, habitat monitoring, scientific exploration, infrastructure protection, health monitoring, and so on. Data collection is a crucial operation in WSNs, where sensor nodes are responsible for collecting all sensory data and delivering them to a Sink node [1]. To understand the physical world in depth, many scientific researchers use the collected observations to reconstruct the environment in cyber space. The reliability of scientific research and decision-making heavily depends on the completeness and accuracy of the environmental observations [2].

However, due to the inherent characteristics of WSNs, data error and data loss are very common in sensor network deployments [3]. They affect the ability of scientists to make meaningful conclusions. For example, Koushanfar and Potkonjak [4] analyzed the sensor data traces across 3 weeks collected from the Intel Berkeley research lab [5]. The results showed that almost 40% of the data was missing and that approximately 8% of the data was faulty. There are a number of reasons for data loss and error. For example, the reasons for data loss include wireless channel instability, inter-channel interference, network congestion, node damage or accidental failure [6], [7]. In addition, the duty-cycle technique used to save energy is also an important reason for data loss [8]. Influenced by some factors during the data-collecting, such as quantization, channel noise, node failure, receiver or base station error, uncertainty of node deployment area, external signal interference, *etc.*,the data collected by some sensor nodes may deviate from the true representation of the physical phenomenon to be measured [9]. Data error occurs when this deviation happens in a sensor node. Data loss and error present significant challenges in

accurately reconstructing the physical world. Therefore, it is necessary to devise an effective method to reconstruct the environmental data from incomplete and erroneous observations to reconstruct the physical world accurately in cyber space.

A great deal of existing work has been devoted to recovering the missing data. K-Nearest-Neighbor (KNN) [10] is a classic interpolation method, which utilizes the adjacent values to estimate the missing data. Delaunay Triangulation (DT) [11] is a typical global refinement method, which treats the gathered data as vertices. DT takes advantage of these vertices and their global errors to build virtual triangles for data interpolation. Multi-channel Singular Spectrum Analysis (MSSA) [12] is a nonparametric and data-adaptive interpolation method based on the embedded lag-covariance matrix, which is a branch of principal component analysis. It is often used in geographic data recovery. The recovery quality of the above methods is generally poor when the data missing rate is high. Moreover, these schemes cannot be applied to handle data error well.

Compressive sensing (CS) is an advanced method to recover the whole data with just a few measurements [13], [14]. Chen *et al.* [15] developed a Multi-Attribute-assistant Compressive Sensing (MACS) algorithm to optimize the recovery accuracy. They proposed a joint sparse decomposition method to find the cross features among multiple attributes based on two real datasets and used the correlation features to jointly recover multi-attribute datasets. Kong *et al.* [18] analyzed the real environmental data and revealed four features of sensory data, such as low-rank, time stability, space similarity and multi-attribute correlation. Based on these observations, they designed an Environmental Space Time Improved Compressive Sensing (ESTI-CS) algorithm with a Multi-Attribute Assistant (MAA) component for estimating the missing data. ESTI-CS calculated the minimal low-rank approximations of the incomplete environmental data matrix, refined the interpolation with spatiotemporal features, and leveraged the strong correlation of multiple attributes from the same dataset for better reconstruction accuracy.

With the rapid progress of sparse representation, matrix completion has been used to recover the missing data in WSNs recently. Cheng*et al.* [16] presented an Efficient Data Collection Approach (EDCA) for data collection in WSNs. EDCA takes advantage of the low-rank feature to achieve both less traffic and high accuracy. To reduce energy consumption of sensors, EDCA randomly chooses the node and time instance to sample data and then uses a matrix completion technique to recover the missing data. However, EDCA only utilizes the low-rank feature of the data matrix. It cannot achieve high accuracy when the data missing rate is high and the empty columns exist in the data matrix. Therefore, Spatiotemporal Compressive Data Collection (STCDG) proposed in [17] made use of both the low-rank and short-term stability features to reduce the amount of traffic and improve the level of recovery accuracy. To avoid the optimization problem involving empty columns, STCDG first removed the empty columns and only recovered the non-empty columns, then filled the empty columns using an optimization technique based on temporal stability. He *et al.* [19] proposed a Data Recovery method with joint Matrix Completion and Sparsity Constraints (DRMCSC). DRMCSC utilized both the low-rank and sparsity features of sensory data to recover the missing data. By utilizing the variable-splitting and penalty techniques, they reformulated the data recovery problem with matrix completion and sparsity constraints as a half-quadratic minimization problem and designed an alternating minimization method to solve it. The algorithms mentioned above only consider the problem of recovering the missing data. However, the data errors and its negative impacts on recovery accuracy of the missing data are not considered.

Since data errors are common problems in data collection, it is imperative to consider their impacts on the data reconstruction. The detection and correction of erroneous data in WSNs have been studied in many works [20], [21]. Ni *et al.* [2] provided a systematically characterized taxonomy of common sensor data faults and presented a systematic way of detecting sensor data faults. The outlier is a pattern that does not match the expected trend in analyzed data. Most outlier detecting algorithms in WSNs exploit the correlation of the sensory data [22]–[24]. Kamal *et al.* [9] proposed a framework for sensor data reliability assessment, Packet-Level Attestation (PLA), which exploits the spatial correlation of data sensed at nearby sensors. PLA is based on the concept of nominating nearby data verifier nodes for each node. A novel sequence-based detection approach, named FIND, for discovering data faults in sensor networks is given in [6]. FIND neither needs a priori knowledge about the underlying distribution of sensed phenomena nor requires costly event injections. In FIND, the faulty nodes are detected based on their violation of the distance monotonicity property in sensing, which is quantified by the metric of ranking differences. Tang *et al.* [25] investigated the impact of outlying sensor readings and broken links on high-fidelity data gathering based on compressive sensing theory. They proposed an approach based on the compressive sensing theory to identify outlying sensor readings, derive the corresponding accurate values, and infer the broken links.

The existing works only separately consider the data recovery problem with missing entries and the data reconstruction problem with erroneous entries, but not consider the mutual influence of missing data and erroneous data in the data reconstruction process. This paper focuses on designing an approach to accurately reconstruct the environmental data in the presence of data loss and data error in WSNs. Our contributions can be summarized as follows:

(1) We propose a novel Data Reconstruction scheme via Temporal Stability guided Matrix Completion, named DRTSMC. To our knowledge, this is the first approach that can simultaneously reconstruct the environmental data

accurately and recognize the sensor nodes that have collected erroneous data.

(2) We first take advantage of the low-rank feature of the raw environmental data to model the data reconstruction problem as a matrix completion model with structural noise. We also introduce the temporal stability to the matrix completion model to further reduce the reconstruction error. Then, we design an algorithm based on the block coordinate descent method and the operator splitting technique to solve the problem.

(3) Finally, we perform simulations with real-world sensory datasets. The results show that the proposed approach significantly outperforms the existing solutions in terms of reconstruction accuracy. In particular, DRTSMC can not only recognize the sensor nodes that have collected erroneous data but also reduce the negative impact of erroneous data on the recovery performance of missing data to improve data reconstruction accuracy.

The rest of this paper is organized as follows: The basic mathematical definitions and theorems are presented in Section II. In Section III, we first introduce the problem definition and system model and then present a data reconstruction approach via temporal stability guided matrix completion. Finally, we evaluate the performance of the proposed DRTSMC through extensive simulations in Section IV, and we conclude the work in Section V.

II. MATHEMATICAL FOUNDATION

In this section, we first give some mathematical definitions and then briefly introduce several theorems that are useful for the subsequent analysis.

Definition 1 (Matrix Norm [26]): Suppose the Singular Value Decomposition (SVD) of matrix $\boldsymbol{X} = (X_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ with rank r: $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the $n_1 \times r$ and $r \times n_2$ matrices, respectively, with orthogonal columns, $\boldsymbol{\Lambda} = diag\{\sigma_i \,|1 \le i \le r\}$ and σ_i is the i-th largest singular value; then,

(1) The Frobenius norm of matrix $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{ij}^2}.$$

(2) The nuclear norm of matrix $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^{r} \sigma_i.$$

(3) The L2,1 norm of matrix $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\|_{2,1} = \sum_{i=1}^{n_1} \left\|X^{(i)}\right\|_2 = \sum_{i=1}^{n_1} \left(\sum_{j=1}^{n_2} X_{ij}^2\right)^{1/2}.$$

Definition 2 (Matrix Shrinkage operator [27]): For any $\tau > 0$, matrix shrinkage operator $D_\tau(\boldsymbol{X})$ is defined as

$$D_\tau(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{S}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^{\mathrm{T}},$$

where $\boldsymbol{S}_\tau(\boldsymbol{\Lambda}) = diag(\{\max(0, \sigma_i - \tau)\,|i = 1, 2, \ldots r\})$.

Theorem 1: For any $\tau, \mu > 0$ and $\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}$, the matrix shrinkage operator $D_{\tau/\mu}(\boldsymbol{Z})$ obeys

$$D_{\tau/\mu}(\boldsymbol{Z}) = \arg\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \left\{\tau \|\boldsymbol{X}\|_* + \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{Z}\|_F^2\right\}. \quad (1)$$

Theorem 2 (Proximal Forward Backward Splitting, PFBS) [28]: Given the following unconstrained convex problem

$$\min_{X \in H} F(\boldsymbol{X}) = F_1(\boldsymbol{X}) + F_2(\boldsymbol{X}), \quad (2)$$

where H is a Hilbert space, both $F_1(\boldsymbol{X})$ and $F_2(\boldsymbol{X})$ are proper lower semi-continuous functions, and $F_2(\boldsymbol{X})$ is smooth with a Lipschitz continuous gradient. Then, the following iterative sequence will converge to a minimizer of convex problem (2):

$$\boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X} \in H} \delta F_1(\boldsymbol{X}) + \frac{1}{2}\left\|\boldsymbol{X} - (\boldsymbol{X}^k - \delta\nabla F_2(\boldsymbol{X}^k))\right\|_F^2, \quad (3)$$

where δ is the step size of iteration and satisfies $0 < \delta < 1/L_f$ and L_f is the Lipschitz continuous gradient of $F_2(\boldsymbol{X})$, i.e., $\exists L_f > 0$, for $\forall \boldsymbol{X}_1, \boldsymbol{X}_2$:

$$\|\nabla F_2(\boldsymbol{X}_2) - \nabla F_2(\boldsymbol{X}_1)\|_F \le L_f \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F. \quad (4)$$

Theorem 3 [29]: For any $\tau, \mu > 0$ and $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_2}$, the function $H(\boldsymbol{X}) = \tau\|\boldsymbol{X}\|_{2,1} + \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{W}\|_F^2$ has a global minimum point $\boldsymbol{X}^* = \mathcal{J}_{\tau/\mu}(\boldsymbol{W})$:

$$\left(\mathcal{J}_{\tau/\mu}(\boldsymbol{W})\right)^{(i)} = \max\left\{\left\|\mathbf{W}^{(i)}\right\|_2 - \tau/\mu, 0\right\} \cdot \mathbf{W}^{(i)} \Big/ \left\|\mathbf{W}^{(i)}\right\|_2 \quad i = 1, 2, \cdots, n_1, \quad (5)$$

where $\left(\mathcal{J}_{\tau/\mu}(\boldsymbol{W})\right)^{(i)}$ represents the i-th row of matrix $\mathcal{J}_{\tau/\mu}(\boldsymbol{W})$ and $\|\cdot\|_2$ denotes the L2-norm of the vector.

III. DATA RECONSTRUCTION VIA TEMPORAL STABILITY GUIDED MATRIX COMPLETION

In this section, we first introduce the problem definition and system model and then present the data reconstruction scheme.

A. PROBLEM DEFINITION

A large number of sensor nodes are deployed in a monitoring area, such as indoors, in a forest or ocean,*etc.*, to collect environmental data to a Sink node. The environmental data reflecting the physical characteristics of the monitoring area can be applied to scientific research. Consider a WSN consisting of one Sink and N sensor nodes, *i.e.*, $v_1, v_2, ..., v_N$. Each sensor node is equipped with one or more environmental sensors for sensing different types of environmental data, such as temperature, humidity, and so on. We employ periodic data collection. The sensor nodes sense and transmit the environmental data to the Sink once every τ time. We call the time interval τ as a time slot. The monitoring time includes T time slots. Therefore, the total amount of data is $N \times T$. These data can be represented by a matrix:

$$\boldsymbol{X} = \begin{bmatrix} x(1,1) & x(1,2) & x(1,3) & \cdots & x(1,T) \\ x(2,1) & x(2,2) & x(2,3) & \cdots & x(2,T) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x(N,1) & x(N,2) & x(N,3) & \cdots & x(N,T) \end{bmatrix} \in \mathbb{R}^{N \times T} \quad (6)$$

where $x(i,j)$ denotes the sensory data of node v_i at time slot j. We define the matrix $\boldsymbol{X}$ as a raw environmental matrix. A complete environmental matrix represents that all environmental data are successfully collected, *i.e.*, there are no missing data.

However, due to the presence of missing data, the Sink node in fact obtains an incomplete matrix. We define this incomplete matrix as a sampling matrix, denoted by $\boldsymbol{R}$. Let $\Omega \subseteq \{1,\cdots,N\} \times \{1,\cdots,T\}$ denote the subscripts set of the observed entries in $\boldsymbol{R}$ and $\mathrm{P}_\Omega(\cdot)$ denote the element-wise projection function as

$$[\mathrm{P}_\Omega(\boldsymbol{R})]_{ij} = \begin{cases} R(i,j) & if\ (i,j) \in \Omega, \\ 0 & otherwise. \end{cases} \tag{7}$$

where $R(i,j)$ is a sampling entry of matrix $\boldsymbol{R}$.

As mentioned in Section I, oss, data error may occur during the data collection. Therefore, there are two types of sampling data. They are raw environmental data and erroneous data. Any sampling data $R(i,j)$ collected by node v_i at time slot j can be expressed as

$$R(i,j) = \begin{cases} X(i,j) & \text{the raw environmental data} \\ F(i,j) & \text{the erroneous data} \end{cases} \tag{8}$$

We can represent the erroneous data as the raw environmental data adding a noise value. Therefore, the erroneous data $F(i,j)$ can be represented as

$$F(i,j) = X(i,j) + Z(i,j), \tag{9}$$

where $Z(i,j)$ is a noise value corresponding to node v_i at time slot j. Therefore, we can define a noise matrix $\boldsymbol{Z}$. Any entry $Z(i,j)$ of matrix $\boldsymbol{Z}$ satisfies: if node v_i collected an erroneous data at time slot j, $Z(i,j) \neq 0$; otherwise, $Z(i,j) = 0$. In the real application scenario, data collected by some sensor nodes are prone to be errors, that is, the entries in some rows of the sampling matrix are prone to be errors. Therefore, some rows in the noise matrix contain nonzero entries, and the remaining rows are all zero entries. We also consider the noise matrix $\boldsymbol{Z}$ as a row-structural noise matrix. Based on the above definitions and analyses, the sampling matrix $\boldsymbol{R}$ can be described by the following equation:

$$\mathrm{P}_\Omega(\boldsymbol{R}) = \mathrm{P}_\Omega(\boldsymbol{X} + \boldsymbol{Z}). \tag{10}$$

For ease of understanding, we give a simple example of data collection, as shown in Fig. 1. A WSN consists of 4 sensor nodes, v_1, v_2, v_3, v_4. They collect data for 6 time slots. The raw environmental matrix $\boldsymbol{X}$ is expressed as

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} \end{bmatrix}. \tag{11}$$

Assume that some of the data collected by node v_2 and v_4 are errors. Data loss and data error are shown in Fig. 1. The corresponding row-structural noise matrix $\boldsymbol{Z}$ is

$$\boldsymbol{Z} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{23} & 0 & z_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & z_{42} & 0 & 0 & z_{45} & 0 \end{bmatrix}. \tag{12}$$

Therefore, the sampling matrix $\boldsymbol{R}$ collected by the WSN is

$$\boldsymbol{R} = \begin{bmatrix} x_{11} & 0 & x_{13} & 0 & x_{15} & 0 \\ x_{21} & x_{22} & x_{23} + z_{23} & x_{24} & x_{25} + z_{25} & 0 \\ 0 & x_{32} & x_{33} & x_{34} & 0 & x_{36} \\ x_{41} & x_{42} + z_{23} & 0 & x_{44} & x_{45} + z_{45} & x_{46} \end{bmatrix}. \tag{13}$$

The key problem to be solved in this paper is how to reconstruct the environmental data matrix from the sampling matrix $\boldsymbol{R}$. In the following section, we will propose a novel Data Reconstruction scheme via Temporal Stability guided Matrix Completion (DRTSMC) to solve the data reconstruction problem.

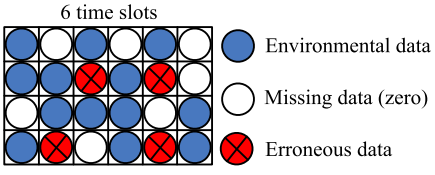


FIGURE 1. The sensory data collected by 4 sensor nodes for 6 time slots.

B. MODEL CONSTRUCTION

According to matrix completion theory, a low-rank or approximately low-rank matrix can be accurately reconstructed from a relatively small number of sampling entries [30], [31]. Based on the analysis of the real environmental datasets gathered by the Intel indoor experiment, GreenOrbs, and Ocean-Sense projects, performed by literature [3], [7], [17], we know that the environmental matrix exhibits the features of low-rank structure and temporal stability. Therefore, the problem of reconstructing the environmental matrix from incomplete and erroneous sensory data can be modeled as a matrix completion problem.

To effectively smooth the structural noise and alleviate its negative impacts on data recovery, DRTSMC introduces the L2,1-norm regularized parameter to the standard matrix completion problem. It applies the L2,1-norm regularized term of the structural noise to the objective function to formulate the data reconstruction as a L2,1-norm regularized matrix completion problem:

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{Z} \in \mathbb{R}^{N \times T}} \|\boldsymbol{X}\|_* + \lambda \|\boldsymbol{Z}\|_{2,1} \\ &s.t.\ \mathrm{P}_\Omega(\boldsymbol{R}) = \mathrm{P}_\Omega(\boldsymbol{X} + \boldsymbol{Z}) \end{aligned} \tag{14}$$

where $\boldsymbol{R} \in \mathbb{R}^{N \times T}$ is a sampling matrix collected by a Sink node and λ is a tunable parameter used to balance the structural noise and the low-rank of the matrix.

Furthermore, the environmental data collected by WSNs usually change slowly over time. In [3], [7], and [17], the researchers explored the feature of temporal stability in

sensory data. They calculated the gap between each pair of adjacent readings for each sensor node, then compared the difference between each pair of adjacent gaps, and found that the sensor readings do not change much in the short term. We introduce a term about temporal stability, $\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2$, to the matrix completion problem (14) to further reduce the recovery error, where $\boldsymbol{S} = Toeplitz(0, 1, -2, 1)$, which denotes the Toeplitz matrix with a central diagonal given by 1, the first upper diagonal given by -2 and the second upper diagonal given by 1. In detail, $\boldsymbol{S}$ can be defined using following equation:

$$\boldsymbol{S} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{T\times T} \tag{15}$$

Finally, we arrive at the following minimization problem:

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{Z}\in\mathbb{R}^{N\times T}} \|\boldsymbol{X}\|_* + \lambda\|\boldsymbol{Z}\|_{2,1} + \frac{\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2 \\ &\quad s.t.\ \mathrm{P}_\Omega(\boldsymbol{R}) = \mathrm{P}_\Omega(\boldsymbol{X}+\boldsymbol{Z}), \end{aligned} \tag{16}$$

where τ is another tunable parameter.

C. MODEL OPTIMIZATION

In this section, we design an efficient optimization algorithm to solve the proposed matrix completion model (16) by employing the block coordinate descent method [26] and the operator splitting technique [28]. Without loss of generality, we reformulate problem (16) as the following equivalent penalty function form:

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{Z}\in\mathbb{R}^{N\times T}} \mu(\|\boldsymbol{X}\|_* + \lambda\|\boldsymbol{Z}\|_{2,1} + \frac{\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2) \\ + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2. \end{aligned} \tag{17}$$

Furthermore, for ease of description, we let:

$$\begin{aligned} L(\boldsymbol{X},\boldsymbol{Z}) = \mu(\|\boldsymbol{X}\|_* + \lambda\|\boldsymbol{Z}\|_{2,1} + \frac{\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2) \\ + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2 \end{aligned} \tag{18}$$

Based on the block coordinate descent method, problem (17) can be solved iteratively as follows:

$$\begin{cases} \boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}\in\mathbb{R}^{N\times T}} L(\boldsymbol{X},\boldsymbol{Z}^k) \\ \boldsymbol{Z}^{k+1} = \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{N\times T}} L(\boldsymbol{X}^{k+1},\boldsymbol{Z}). \end{cases} \tag{19}$$

1) For sub-problem 1:

$$\boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}\in\mathbb{R}^{N\times T}} L(\boldsymbol{X},\boldsymbol{Z}^k) \tag{20}$$

We have:

$$\begin{aligned} \boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}\in\mathbb{R}^{N\times T}} \mu(\|\boldsymbol{X}\|_* + \frac{\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2) \\ + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2 \end{aligned} \tag{21}$$

Furthermore, let:

$$F_1(\boldsymbol{X}) = \mu\|\boldsymbol{X}\|_*, \tag{22}$$

$$F_2(\boldsymbol{X}) = \frac{\mu\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2 + \frac{1}{2}\left\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z}^k)\right\|_F^2. \tag{23}$$

We can see that both $F_1(\boldsymbol{X})$ and $F_2(\boldsymbol{X})$ are lower semi-continuous convex functions, and $F_2(\boldsymbol{X})$ is differentiable on $\mathbb{R}^{N\times T}$. According to the **Theorem 2**, we have

$$\begin{aligned} &\boldsymbol{X}^{k+1} \\ &= \arg\min_{\boldsymbol{X}\in\mathbb{R}^{N\times T}} \begin{cases} \mu\delta_X\|\boldsymbol{X}\|_* \\ +\frac{1}{2}\left\|\boldsymbol{X} - \left(\begin{matrix}\boldsymbol{X}^k - \delta_X\mu\tau\boldsymbol{X}^k\boldsymbol{S}^T\boldsymbol{S} \\ +\delta_X\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}^k-\boldsymbol{Z}^k)\end{matrix}\right)\right\|_F^2 \end{cases} \end{aligned} \tag{24}$$

Let $\boldsymbol{U}^k = \boldsymbol{X}^k - \delta_X\mu\tau\boldsymbol{X}^k\boldsymbol{S}^T\boldsymbol{S} + \delta_X\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}^k-\boldsymbol{Z}^k)$; then, Eq. (18) can be simplified to

$$\boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}\in\mathbb{R}^{N\times T}} \mu\delta_X\|\boldsymbol{X}\|_* + \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{U}^k\right\|_F^2. \tag{25}$$

According to **Theorem 1**, $\boldsymbol{X}^{k+1}$ can be expressed as

$$\boldsymbol{X}^{k+1} = D_{\mu\delta_X}(\boldsymbol{U}^k), \tag{26}$$

where δ_X is a step size. Specifically, since we have the following inequation:

$$\begin{aligned} &\|\nabla F_2(X_1) - \nabla F_2(X_2)\|_F^2 \\ &= \left\|\mathrm{P}_\Omega(\boldsymbol{X}_1-\boldsymbol{X}_2) + \mu\tau(\boldsymbol{X}_1-\boldsymbol{X}_2)\boldsymbol{S}^T\boldsymbol{S}\right\|_F^2 \\ &\le 2\|\mathrm{P}_\Omega(\boldsymbol{X}_1-\boldsymbol{X}_2)\|_F^2 + 2\mu^2\tau^2\sigma_1^2(\boldsymbol{S}^T\boldsymbol{S})\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_F^2 \\ &\le \left(2+2\mu^2\tau^2\sigma_1^2(\boldsymbol{S}^T\boldsymbol{S})\right)\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_F^2 \end{aligned} \tag{27}$$

where $\sigma_1(\boldsymbol{S}^T\boldsymbol{S})$ is the largest singular value of matrix $\boldsymbol{S}^T\boldsymbol{S}$, we can set the Lipschitz continuous gradient of $F_2(\boldsymbol{X})$ as

$$L_f = \sqrt{2+2\mu^2\tau^2\sigma_1^2(\boldsymbol{S}^T\boldsymbol{S})} \tag{28}$$

In this paper, according to **Theorem 2,**we set

$$\delta_X = \frac{1}{\sqrt{2+2\mu^2\tau^2\sigma_1^2(\boldsymbol{S}^T\boldsymbol{S})}} \tag{29}$$

Therefore, Sub-problem 1 can be solved by the following iterative method:

$$\begin{cases} \boldsymbol{U}^k = \boldsymbol{X}^k - \delta_X\mu\tau\boldsymbol{X}^k\boldsymbol{S}^T\boldsymbol{S} + \delta_X\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}^k-\boldsymbol{Z}^k) \\ \boldsymbol{X}^{k+1} = D_{\mu\delta_X}(\boldsymbol{U}^k). \end{cases} \tag{30}$$

2) For Sub-problem 2:

$$\boldsymbol{Z}^{k+1} = \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{N\times T}} L(\boldsymbol{X}^{k+1},\boldsymbol{Z}) \tag{31}$$

We have:

$$\begin{aligned}\boldsymbol{Z}^{k+1} &= \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{N\times T}} \mu\lambda \|\boldsymbol{Z}\|_{2,1} + \frac{1}{2}\left\|\mathrm{P}_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^{k+1}-\boldsymbol{Z})\right\|_F^2 \\ &= \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{N\times T}} \mu\lambda\delta_Z \|\boldsymbol{Z}\|_{2,1} + \frac{1}{2}\left\|\boldsymbol{Z}-\boldsymbol{V}^k\right\|_F^2, \end{aligned} \tag{32}$$

where $\boldsymbol{V}^k = \boldsymbol{Z}^k + \delta_Z \mathrm{P}_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k)$ and δ_Z satisfy the Lipschitz continuous gradient, which is determined as follows:

$$\begin{aligned}&\left\|\mathrm{P}_{\Omega}(\boldsymbol{X}_1+\boldsymbol{Z}^k-\boldsymbol{R})-\mathrm{P}_{\Omega}(\boldsymbol{X}_2+\boldsymbol{Z}^k-\boldsymbol{R})\right\|_{\boldsymbol{F}}^2 \\ &\quad= \|\mathrm{P}_{\Omega}(\boldsymbol{X}_1-\boldsymbol{X}_2)\|_{\boldsymbol{F}}^2 \\ &\quad\le \|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\boldsymbol{F}}^2 \end{aligned} \tag{33}$$

i.e., the Lipschitz constant $L_f = 1$. The step size of iteration satisfies $0 < \delta_Z < 1$.

According to **Theorem 3**, $\boldsymbol{Z}^{k+1}$ can be solved as follows:

$$\boldsymbol{Z}^{k+1} = \mathcal{J}_{\mu\lambda\delta_Z}(\boldsymbol{V}^k). \tag{34}$$

Finally, Sub-problem 2 can be solved by the following iterative method, i.e.,

$$\begin{cases}\boldsymbol{V}^k = \boldsymbol{Z}^k + \delta_Z \mathrm{P}_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k) \\ \boldsymbol{Z}^{k+1} = \mathcal{J}_{\mu\lambda\delta_Z}(\boldsymbol{V}^k)\end{cases} \tag{35}$$

D. ALGORITHM IMPLEMENTATION

Based on the above analyses, we can get the iterative solution of the minimization problem (13). The data recovery algorithm based on the operator splitting technique is shown in **Algorithm 1**. The input of **Algorithm 1** is the sampling matrix $\boldsymbol{R}$ collected by the WSN, the maximum number of iterations $\boldsymbol{R}$, and various parameters, such as μ, λ, τ, *etc.* The output of**Algorithm 1** is recovered data matrix $\boldsymbol{X}_{opt}$ and recovered noise matrix $\boldsymbol{Z}_{opt}$. In **Algorithm 1**, we first set the initial matrix $\boldsymbol{X}^0$ and $\boldsymbol{Z}^0$ as 0 (line 1). Sub-problems 1 and 2 are solved in lines 3-4 and lines 5-6, respectively.

Algorithm 1 Data Recovery Algorithm Based on the Block Coordinate Descent Method

Input: sampling matrix $\boldsymbol{R}$, μ, λ, τ, maximum number of iterations *Max*

Output: $\boldsymbol{X}_{opt}\boldsymbol{Z}_{opt}$

1) Initialization $\delta_Z = \frac{1}{2}$,

$$\delta_X = \frac{1}{\sqrt{2+2\mu^2\tau^2\sigma_1^2(\boldsymbol{S}^T\boldsymbol{S})}}, \quad \boldsymbol{X}^0 = 0,\ \boldsymbol{Z}^0 = 0;$$

2) FOR $k = 0$ to *Max*
3) $\quad \boldsymbol{U}^k = \boldsymbol{X}^k - \delta_X\mu\tau\boldsymbol{X}^k\boldsymbol{S}^T\boldsymbol{S} + \delta_X\mathrm{P}_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^k-\boldsymbol{Z}^k)$;
4) $\quad \boldsymbol{X}^{k+1} = D_{\mu\delta_X}(\boldsymbol{U}^k)$;
5) $\quad \boldsymbol{V}^k = \boldsymbol{Z}^k + \delta_Z\mathrm{P}_{\Omega}(\boldsymbol{R}-\boldsymbol{X}^{k+1}-\boldsymbol{Z}^k)$;
6) $\quad \boldsymbol{Z}^{k+1} = \mathcal{J}_{\mu\lambda\delta_Z}(\boldsymbol{V}^k)$
7) END FOR
8) RETURN $\boldsymbol{X}_{opt} \leftarrow \boldsymbol{X}^{Max+1}$, $\boldsymbol{Z}_{opt} \leftarrow \boldsymbol{Z}^{Max+1}$

Using matrix $\boldsymbol{X}_{opt}$ and $\boldsymbol{Z}_{opt}$, DRTSMC reconstructs the environment matrix $\boldsymbol{X}_{rec}$ through the following two steps:

Step 1: We first recover the missing data by inserting the entries of the recovered data matrix $\boldsymbol{X}_{opt}$ in the corresponding missing point. Any entry in $\boldsymbol{X}_{rec}$ satisfies

$$X_{rec}(i,j) = \begin{cases}R(i,j) & (i,j)\in\Omega \\ X_{opt}(i,j) & otherwise.\end{cases} \tag{36}$$

Step 2: The faulty nodes can be recognized through analyzing the recovered noise matrix $\boldsymbol{Z}_{opt}$. In matrix $\boldsymbol{Z}_{opt}$, the rows, whose entries are all zeros, correspond to the sensor nodes without erroneous data. Otherwise, the rows with nonzero entries correspond to the faulty sensor nodes. After recognizing the faulty nodes, we can replace the rows containing erroneous data in the reconstructed matrix $\boldsymbol{X}_{rec}$ with the corresponding data rows of matrix $\boldsymbol{X}_{opt}$, *i.e.,* any raw in $\boldsymbol{X}_{rec}$ satisfies

$$\begin{aligned}&\boldsymbol{X}_{rec}^{(i)} \\ &\quad= \begin{cases}\boldsymbol{X}_{rec}^{(i)} & Z_{opt}(i,j) = 0 \forall Z_{opt}(i,j)\in\boldsymbol{Z}_{opt}^{(i)}, \\ & \qquad\qquad j = 1,\ldots T, \\ \boldsymbol{X}_{opt}^{(i)} & otherwise\end{cases}\end{aligned} \tag{37}$$

where $\boldsymbol{X}_{rec}^{(i)}$ and $\boldsymbol{X}_{opt}^{(i)}$ represent the i-th row of matrix $\boldsymbol{X}_{rec}$ and $\boldsymbol{X}_{opt}$, respectively.

As the example given in Subsection *A*, the method of reconstructing the environmental data matrix $\boldsymbol{X}_{rec}$ can be illustrated by Fig. 2. The sampling matrix is shown in Fig. 2 (a). Fig. 2 (b) and (c) are matrix $\boldsymbol{X}_{opt}$ and $\boldsymbol{Z}_{opt}$, respectively. The reconstructed environmental matrix $\boldsymbol{X}_{rec}$ is shown in Fig. 2 (d).

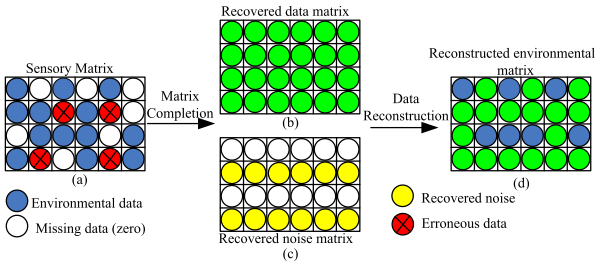


FIGURE 2. The method of reconstructing the environmental matrix.

E. CONVERGENCE ANALYSIS

Theoretically, for the jointly convex problem with the separable non-smooth terms, Tseng [32] has demonstrated that the block coordinate descent method is guaranteed to converge to a global optimum, as long as all sub-problems are solvable. In our model, it is obvious that the model's non-smooth parts, i.e., both $\mu\|\boldsymbol{X}\|_*$ and $\lambda\|\boldsymbol{Z}\|_{2,1}$, are separable, and based on Theorem 2 we can see that both the two sub-problems are solvable. Furthermore, we can also prove that the objective

function of this model is jointly convex for $\boldsymbol{X}$ and $\boldsymbol{Z}$ by using **Proposition 1.** Therefore, based on this fact, we can easily draw a conclusion that our proposed optimization algorithm also has the provable convergence.

Proposition 1: The model proposed in our work:

$$L(\boldsymbol{X},\boldsymbol{Z}) = \mu(\|\boldsymbol{X}\|_* + \lambda\|\boldsymbol{Z}\|_{2,1} + \frac{\tau}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2) + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2 \quad (38)$$

is a jointly convex model, where $\boldsymbol{S} \in \mathbb{R}^{T\times T}$.

Proof: Obviously, the domain of this model $\left\{(\boldsymbol{X},\boldsymbol{Z})|\ \boldsymbol{X},\boldsymbol{Z} \in \mathbb{R}^{N\times T}\right\}$ is a convex set.Meanwhile, from [26], we know that the nuclear norm and L2,1-norm are convex, and a nonnegative weighted sum of convex functions is also convex. Therefore, to prove that $L(\boldsymbol{X},\boldsymbol{Z})$ is jointly convex, we only need to prove that**:**

$$\frac{\tau\mu}{2}\left\|\boldsymbol{X}\boldsymbol{S}^T\right\|_F^2 + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2 \quad (39)$$

is jointly convex.

For ease of description, formula (39) can be simplified as:

$$G(\boldsymbol{X},\boldsymbol{Z}) = \frac{k}{2}\|\boldsymbol{X}\boldsymbol{M}\|_F^2 + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{R}-\boldsymbol{X}-\boldsymbol{Z})\|_F^2 \quad (40)$$

where $k{=}\tau\mu$, $\boldsymbol{M} = \boldsymbol{S}^T$.

Next, we will demonstrate that $G(\boldsymbol{X},\boldsymbol{Z})$ is jointly convex by proving that it obeys the following first-order conditions of convex function for any $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^{N\times T}$ [33]:

$$G(\boldsymbol{X}_2,\boldsymbol{Z}_2) \geq G(\boldsymbol{X}_1,\boldsymbol{Z}_1) + \langle\nabla G(\boldsymbol{X}_1,\boldsymbol{Z}_1), \begin{bmatrix}\boldsymbol{X}_2-\boldsymbol{X}_1\\ \boldsymbol{Z}_2-\boldsymbol{Z}_1\end{bmatrix}\rangle \quad (41)$$

Obviously, for $G(\boldsymbol{X},\boldsymbol{Z})$, we have:

$$\nabla G(\boldsymbol{X},\boldsymbol{Z}) = \begin{bmatrix}k\boldsymbol{X}\boldsymbol{M}\boldsymbol{M}^T\\ 0_{N\times T}\end{bmatrix} - \begin{bmatrix}\mathrm{P}_\Omega(\boldsymbol{X}+\boldsymbol{Z}-\boldsymbol{R})\\ \mathrm{P}_\Omega(\boldsymbol{X}+\boldsymbol{Z}-\boldsymbol{R})\end{bmatrix} \quad (42)$$

Then:

$$\begin{aligned}
&G(\boldsymbol{X}_2,\boldsymbol{Z}_2) - G(\boldsymbol{X}_1,\boldsymbol{Z}_1) - \langle\nabla G(\boldsymbol{X}_1,\boldsymbol{Z}_1), \begin{bmatrix}\boldsymbol{X}_2-\boldsymbol{X}_1\\ \boldsymbol{Z}_2-\boldsymbol{Z}_1\end{bmatrix}\rangle\\
&= \frac{k}{2}\|\boldsymbol{X}_2\mathbf{M}\|_F^2 - \frac{k}{2}\|\boldsymbol{X}_1\mathbf{M}\|_F^2 + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{R})\|_F^2\\
&\quad - \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{X}_1+\boldsymbol{Z}_1-\boldsymbol{R})\|_F^2\\
&\quad - \mathrm{tr}\Big(\mathrm{P}_\Omega\Big((\boldsymbol{X}_2-\boldsymbol{X}_1)^T(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{R})\\
&\quad + (\boldsymbol{Z}_2-\boldsymbol{Z}_1)^T(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{R})\Big)\Big)\\
&\quad - k\cdot\mathrm{tr}\Big((\boldsymbol{X}_2-\boldsymbol{X}_1)^T(\boldsymbol{X}_1\mathbf{M}\mathbf{M}^T)\Big)\\
&= \frac{k}{2}\cdot\mathrm{tr}\Big((\boldsymbol{X}_2\mathbf{M}-\boldsymbol{X}_1\boldsymbol{M})^T(\boldsymbol{X}_2\mathbf{M}+\boldsymbol{X}_1\mathbf{M})^T\Big)\\
&\quad - k\cdot\mathrm{tr}\Big((\boldsymbol{X}_2\mathbf{M}-\boldsymbol{X}_1\boldsymbol{M})^T(\boldsymbol{X}_1\mathbf{M})\Big)\\
&\quad + \frac{1}{2}\mathrm{tr}\Big(\mathrm{P}_\Omega\Big((\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{X}_1-\boldsymbol{Z}_1)^T\\
&\quad \times(\boldsymbol{X}_2+\boldsymbol{Z}_2+\boldsymbol{X}_1+\boldsymbol{Z}_1-2\mathbf{R})\Big)\Big)\\
&\quad - \mathrm{tr}\Big(\mathrm{P}_\Omega\Big((\boldsymbol{X}_2-\boldsymbol{X}_1)^T(\boldsymbol{X}_1+\boldsymbol{Z}_1-\mathbf{R})\\
&\quad + (\boldsymbol{Z}_2-\boldsymbol{Z}_1)^T(\boldsymbol{X}_1+\boldsymbol{Z}_1-\mathbf{R})\Big)\Big)\\
&= \frac{k}{2}\cdot\mathrm{tr}\Big((\boldsymbol{X}_2\mathbf{M}-\boldsymbol{X}_1\boldsymbol{M})^T(\boldsymbol{X}_2\mathbf{M}+\boldsymbol{X}_1\mathbf{M}-2\boldsymbol{X}_1\mathbf{M})\Big)\\
&\quad + \frac{1}{2}\mathrm{tr}\Big(\mathrm{P}_\Omega\Big((\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{X}_1-\boldsymbol{Z}_1)^T(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{X}_1-\boldsymbol{Z}_1)\Big)\Big)\\
&= \frac{k}{2}\cdot\mathrm{tr}\Big(\mathbf{M}^T(\boldsymbol{X}_2-\boldsymbol{X}_1)^T(\boldsymbol{X}_2-\boldsymbol{X}_1)\mathbf{M}\Big)\\
&\quad + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{X}_1-\boldsymbol{Z}_1)\|_F^2\\
&= \frac{k}{2}\|(\boldsymbol{X}_2-\boldsymbol{X}_1)\mathbf{M}\|_F^2 + \frac{1}{2}\|\mathrm{P}_\Omega(\boldsymbol{X}_2+\boldsymbol{Z}_2-\boldsymbol{X}_1-\boldsymbol{Z}_1)\|_F^2 \geq 0
\end{aligned} \quad (43)$$

Therefore, we draw a conclusion that:

$$G(\boldsymbol{X}_2,\boldsymbol{Z}_2) \geq G(\boldsymbol{X}_1,\boldsymbol{Z}_1) + \langle\nabla G(\boldsymbol{X}_1,\boldsymbol{Z}_1), \begin{bmatrix}\boldsymbol{X}_2-\boldsymbol{X}_1\\ \boldsymbol{Z}_2-\boldsymbol{Z}_1\end{bmatrix}\rangle \quad (44)$$

That is, the function $G(\boldsymbol{X},\boldsymbol{Z})$ is jointly convex, which means that our proposed model is jointly convex. ■

IV. PERFORMANCE EVALUATIONS

To evaluate the performance of our scheme, we perform the extensive simulations driven by real-world environmental datasets and compare the proposed DRTSMC with the state-of-the-art STCDG [17], DRMCSC [19] and RPCA [34] in this section.

A. EXPERIMENTAL ENVIRONMENT

We use the real-world environmental datasets from the Intel Indoor project [5] to perform the simulations. In the Intel indoor experiment, there are 54 Mica2Dot nodes placed in a 40m × 30m room. Each node reports once every 30 seconds. The sensory data include temperature, light, and humidity. We select temperature data as the raw experimental dataset. The experimental dataset contains the data collected by 52 sensor nodes in 300 consecutive time slots, i.e., $N = 52$, $T = 300$. The temperature range is [15.8195, 27.7836].

Let n be the number of missing entries in the sampling matrix; then, the data missing rate p_n can be expressed as $p_n = n/N \times T$. We define the data sampling rate P_s as $p_s = 1 - p_n$. The ratio of the number of faulty sensor nodes to the total number of sensor nodes is called the rate of faulty sensor nodes. Let m be the number of faulty sensor nodes. Then, the rate of faulty sensor nodes $p_m = m/N$.

To evaluate the reconstruction performance, we generate a data matrix with random data missing as well as structural data error from the raw experimental data matrix. We denote the raw data matrix as $\boldsymbol{X}_{N\times T}$. From the raw data, we generate the synthesized experimental data, denoted as $\boldsymbol{R}_{N\times T}$. The synthesized data $\boldsymbol{R}_{N\times T}$ are generated through the following two steps:

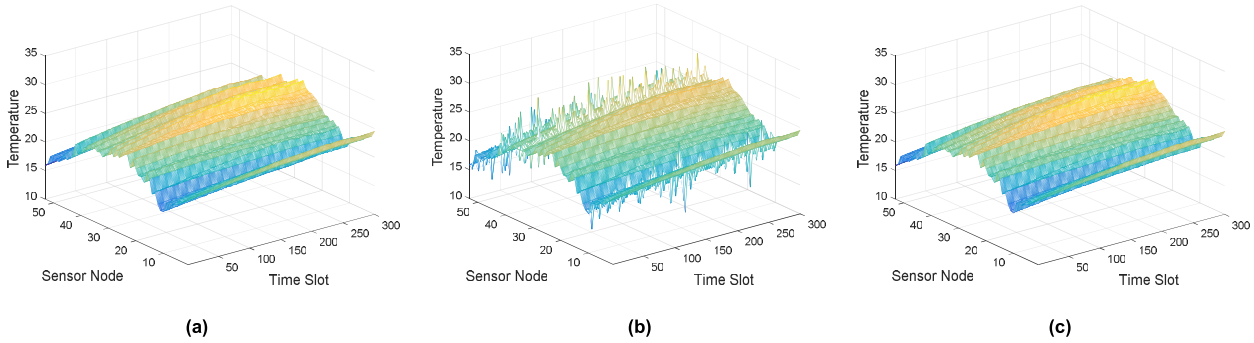


FIGURE 3. The comparison of data matrix. (a) The raw environmental data. (b) Data reconstructed by DRMCSC. (c) Data reconstructed by DRTSMC.

Step 1: According to the data sampling rate p_s, we determine the random subscripts set Ω of the observed entries. We then sample the entries from the raw matrix $\boldsymbol{X}_{N\times T}$ according to the subscripts set Ω. After this step, the synthesized data $\boldsymbol{R}$ can be expressed as follows.

$$R_{ij} = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & otherwise, \end{cases} \tag{45}$$

Step 2: Based on the given rate of faulty sensor nodes p_m, we can determine the number of faulty sensor nodes, *i.e.*, $m = \lfloor p_m \times N \rfloor$. To generate structural erroneous data, we randomly select m rows from $\boldsymbol{R}$, in which 50% of nonzero entries are set as erroneous data by adding randomly generated noise. Denote the erroneous data location set as. The synthesized sampling data can be expressed as

$$R_{ij} = \begin{cases} R_{ij} + Z_{ij} & (i,j) \in \Omega \\ R_{ij} & otherwise, \end{cases} \tag{46}$$

where Z_{ij} is the generated noise at location (i, j) following a zero-mean normal distribution with variance δ^2, i.e., $Z_{ij} \sim N(0, \delta^2)$. In this simulation, we set $\delta^2 = 4$.

After the above two steps, the synthesized data matrix is obtained. We then use the synthesized data matrix $\boldsymbol{R}_{N\times T}$ as the sampling matrix for reconstructing the environmental data matrix. Finally, we verify the performance of our proposed DRTSMC by comparing the reconstruction data matrix with the raw data matrix$\boldsymbol{X}_{N\times T}$.

In **Algorithm 1**, the theoretical research for adaptive setting of tunable parameters μ, λ and τ has not yet been carried out. In the simulation process, we cross-validate the tunable parameters λ and τ based on the prior knowledge of the problem that we are dealing with. At the same time, in order to speed up the convergence of the algorithm, the initial value of the parameter μ is set to the L2 norm of the sampling matrix, and then, it iterates to 0.01 at the rate of 0.25.

B. DEFINITIONS OF PERFORMANCE PARAMETERS

To measure the performance, some definitions of performance parameters are given in this subsection.

Definition 3. Recovery Error of Missing Data (ε^{miss}) is a metric for measuring the error in the recovery of the missing entries in the matrix:

$$\varepsilon^{miss} = \frac{\sqrt{\sum_{i,j:(i,j)\in\Omega^M}(X(i,j) - X_{rec}(i,j))^2}}{\sqrt{\sum_{i,j:(i,j)\in\Omega^M}(X(i,j))^2}}, \tag{47}$$

where Ω^M denotes the subscripts set of the missing-data points in the sampling matrix. The ε^{miss} reflects the ability of the algorithm to recover the missing data.

Definition 4: Recognition Rate of Faulty Sensor Node (r^{node}) is a metric to measure the ability to recognize the faulty sensor node:

$$r^{node} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{48}$$

$$precision = \frac{m\_true}{m\_all} \tag{49}$$

$$recall = \frac{m\_true}{m}, \tag{50}$$

where *m_all* represents the number of faulty sensor nodes recognized by the proposed method, and *m_true* denotes the number of true faulty nodes among them, m is the actual number of faulty sensor nodes, i.e., $m = \lfloor p_m \times N \rfloor$

Definition 5: Reconstruction Error of Erroneous Row (ε^{row}). As described in Section III, we replace the erroneous data rows of matrix $\boldsymbol{X}_{rec}$ with the corresponding data rows in matrix $\boldsymbol{X}_{opt}$. The error caused by this replacement is defined as reconstruction error of erroneous row. It can be expressed as

$$\varepsilon^{row} = \frac{\sqrt{\sum_{i,j:i\in\Theta,j\in[1,\ldots,T]}(X(i,j) - X_{rec}(i,j))^2}}{\sqrt{\sum_{i,j:i\in\Theta,j\in[1,\ldots,T]}(X(i,j))^2}} \tag{51}$$

where Θ is the set of faulty sensor nodes. The ε^{row} reflects the ability of the algorithm to recover erroneous data.

C. PERFORMANCE COMPARISONS

In this section, we compare the proposed DRTSMC with the state-of-the-art STCDG [17], DRMCSC [19] and RPCA [34]. We report an average of 30 random runs.

We first give a straightforward comparison of the data matrix reconstructed by the different algorithms. In Fig. 3, we show the data reconstruction performance in the case

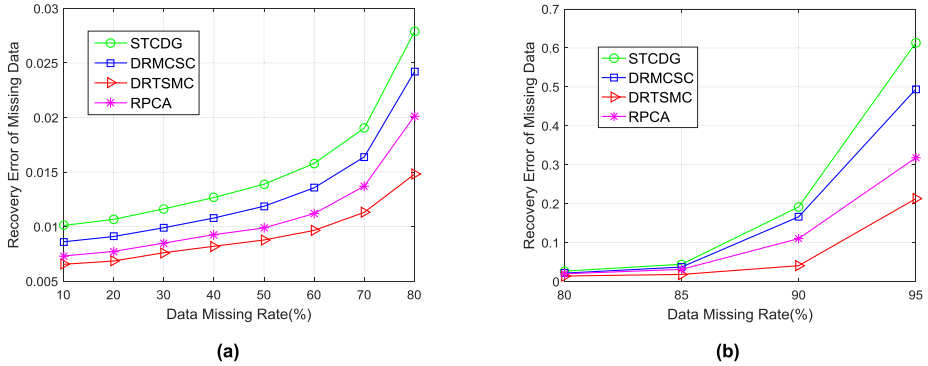


FIGURE 4. Recovery error of missing data under different missing rates. (a) The data missing rate from 10% to 80%. (b) The data missing rate from 80% to 95%.

that the faulty rate $p_m = 10\%$ and the data missing rate $p_n = 30\%$. The 3D image of the raw environmental matrix is shown in Fig. 3(a). Fig. 3 (b) and (c) show the 3D images of the data matrices reconstructed by DRMCSC and DRTSMC respectively. It can be seen intuitively that the data matrix reconstructed by our method is closer to the raw data matrix, that is, the missing data can be recovered and the erroneous data rows can be corrected effectively. In addition, through Fig. 3(b), we also find that the erroneous data have a significant impact on the data reconstruction accuracy of DRMCSC.

Fig.4 shows the data recovery performance of several algorithms in the case of the same faulty rate and different data missing rates. In Fig. 4, the faulty rate p_m is set to be 20%, and the data missing rate p_n ranges from 10% to 95%. To facilitate observation, the comparison results are shown by two figures. As shown in Fig. 4, the recovery error of DRTSMC is always lower than that of other algorithms, i.e., our algorithm shows the best recovery performance. When the data missing rate is low, the recovery errors of all algorithms are relatively small. The recovery error increases with the data missing rate. However, even when 90% of the data have been lost, the recovery error of DRTSMC is still less than 5%, while the recovery error of RPCA is more than 10%, and the other two algorithms are close to 20%. When the data missing rate exceeds 90%, the recovery error increases dramatically. In addition, when the data missing rate is relatively high, DRTSMC has a more obvious advantage over other algorithms.

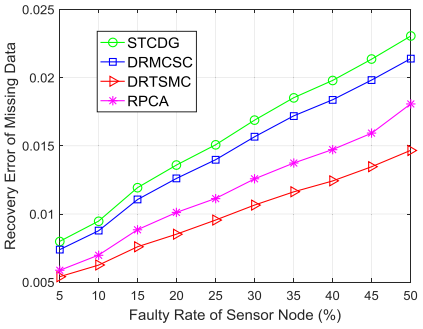


FIGURE 5. Recovery error of missing data under different faulty rates.

Fig.5 depicts the recovery error of missing data under different faulty rates. In Fig. 5, the data missing rate p_n is fixed at 50%. We increase the rate of faulty sensor nodes from 5% to 50%. In general, the recovery error of the four algorithms increases with the rate of faulty sensor nodes. It reveals the fact that the erroneous data have a significant impact on the recovery performance of the missing data. However, DRTSMC is much better than other algorithms. With the increase in the faulty rate, the performance of DRTSMC is increasingly superior to the other algorithms.

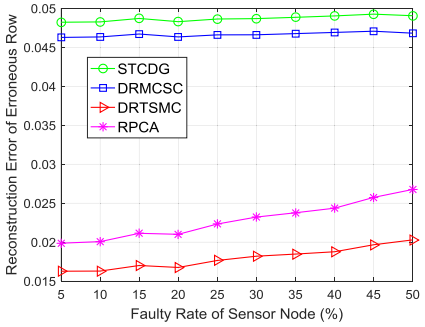


FIGURE 6. Reconstruction error of erroneous rows under different faulty rates.

The performance comparison of reconstruction error of erroneous rows is shown in Fig. 6. Because the STCDG and DRMCSC cannot recognize the faulty sensor nodes, the reconstruction error is larger. As described in [34], the RPCA algorithm works well for the outlier noise with uniform or approximately uniform distribution, while the noise we suffered from is the row-structural noise, which does not obey uniform or approximately uniform distribution. Therefore, the reconstruction error of RPCA is larger than DRTSMC. The performance of DRTSMC is better than other algorithms, which is also clearly shown in Fig. 3. In addition, as the faulty rate increases, the reconstruction error of erroneous rows increases slightly.

We now explore the faulty-node recognition ability of DRTSMC. First, we study the recognition rate of faulty sensor nodes under different data missing rates. In Fig. 7, the faulty rates are set to 40% and 60%. The data missing rate p_n ranges from 10% to 90%. When the data missing rate increases to 60%, the recognition rate of faulty sensor nodes is still close

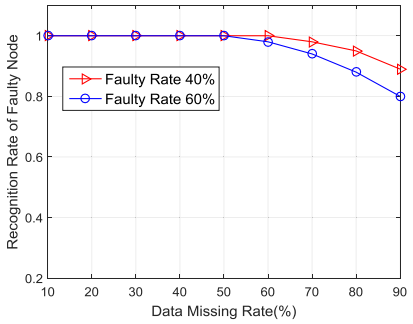


FIGURE 7. Recognition rate under different missing rates.

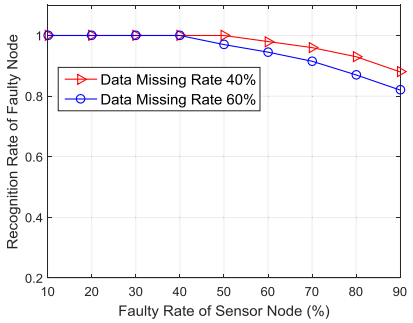


FIGURE 8. Recognition rate under different faulty rates.

to 100%. It indicates that DRTSMC has a strong ability to recognize the faulty sensor nodes. With a further increase in the data missing rate, the recognition rate decreased slightly. Even if the data missing rate reaches 90%, the recognition rate is still more than 80%. Then, we analyze the recognition ability under different faulty rates. Fig. 8 shows the evolution of the recognition rate of DRTSMC when the data missing rate is 40% and 60%, and the faulty rate ranges from 10% to 90%. In Fig. 8, even if the faulty rate rises to 60%, the recognition rate can still be close to 100%. When the faulty rate is further increased, the recognition rate decreases slightly. From the above two figures, we can conclude that DRTSMC can completely recognize the faulty sensor nodes when the faulty rate and data missing rate are not too high, i.e., our algorithm has strong recognition ability of faulty sensor nodes.

In summary, DRTSMC outperforms STCDG, DRMCSC and RPCA in terms of reconstruction accuracy. DRTSMC not only can exactly recognize the sensor nodes that collected erroneous sensory data but can also effectively reconstruct the environmental data.

V. CONCLUSION

Aiming at the problem of data loss and error in the process of data collection in WSNs, this paper proposes a data reconstruction scheme based on matrix completion and temporal stability. The scheme applies matrix completion to fully exploit the low-rank and temporal stability features of environmental data to reconstruct environmental data. We formulate the data reconstruction problem as a L2,1-norm regularized matrix completion model. We also design an algorithm based on the block coordinate descent method and the operator splitting technique to realize the reconstruction of environmental data.

We have performed extensive simulations with real-world sensory datasets. The simulation results demonstrate that our DRMCSC can achieve very good reconstruction performance when data loss and error exist simultaneously. Most importantly, our DRMCSC can recognize the faulty sensor nodes and reduce the negative impact of erroneous data on data reconstruction to improve data reconstruction performance.

REFERENCES

[1] X. Wu, Y. Xiong, P. Yang, S. Wan, and W. Huang, "Sparsest random scheduling for compressive data gathering in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5867–5877, Oct. 2014.

[2] K. Ni *et al.*, "Sensor network data fault types," *ACM Trans. Sensor Netw.*, vol. 5, no. 3, pp. 1–29, 2009.

[3] 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.

[4] F. Koushanfar and M. Potkonjak, "Markov chain-based models for missing and faulty data in MICA2 sensor motes," in *Proc. 4th IEEE Conf. Sensors*, Irvine, CA, USA, Oct./Nov. 2005, pp. 1430–1434.

[5] *Intel Indoor Data Trace*. Accessed: Mar. 14, 2018. [Online]. Available: http://db.lcs.mit.edu/labdata/labdata.html

[6] S. Guo, H. Zhang, Z. Zhong, J. Chen, Q. Cao, and T. He, "Detecting faulty nodes with data errors for wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 10, no. 3, p. 40, 2014.

[7] X. Liu, J. Li, Z. Dong, and F. Xiong, "Joint design of energy-efficient clustering and data recovery for wireless sensor networks," *IEEE Access*, vol. 5, pp. 3646–3656, 2017.

[8] L. Kong *et al.*, "Resource-efficient data gathering in sensor networks for environment reconstruction," *Comput. J.*, vol. 58, no. 6, pp. 1330–1343, 2014.

[9] A. R. M. Kamal, C. Bleakley, and S. Dobson, "Packet-Level Attestation (PLA): A framework for in-network sensor data reliability," *ACM Trans. Sensor Netw.*, vol. 9, no. 2, p. 19, 2013.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[11] L. Kong, D. Jiang, and M.-Y. Wu, "Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction," in *Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst.*, Jun. 2010, pp. 179–188.

[12] H. Zhu, Y. Zhu, M. Li, and L. M. Ni, "SEER: Metropolitan-scale traffic perception based on lossy sensory data," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 217–225.

[13] Z. Zou, Y. Bao, H. Li, B. F. Spencer, and J. Ou, "Embedding compressive sensing-based data loss recovery algorithm into wireless smart sensors for structural health monitoring," *IEEE Sensors J.*, vol. 15, no. 2, pp. 797–808, Feb. 2015.

[14] Z. Chen, G. Yang, L. Chen, and J. Xu, "Constructing maximum-lifetime data-gathering tree in WSNs based on compressed sensing," *Int. J. Distrib. Sensor Netw.*, vol. 12, no. 5, p. 2313064, 2016.

[15] G. Chen *et al.*, "Multiple attributes-based data recovery in wireless sensor networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 103–108.

[16] J. Cheng *et al.*, "Efficient data collection with sampling in WSNs: making use of matrix completion techniques," in *Proc. GLOBECOM*, Miami, FL, USA, Dec. 2010, pp. 1–5.

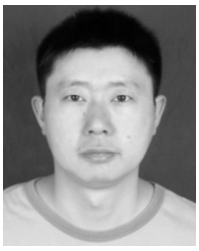
[17] J. Cheng, Q. Ye, H. Jiang, D. Wang, and C. Wang, "STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 850–861, Feb. 2013.

[18] L. Kong *et al.*, "Data loss and reconstruction in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2818–2828, Nov. 2014.

[19] J. He, G. Sun, Y. Zhang, and Z. Wang, "Data recovery in wireless sensor networks with joint matrix completion and sparsity constraints," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2230–2233, Dec. 2015.
[20] A. Mahapatro and P. M. Khilar, "Fault diagnosis in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2000–2026, 4th Quart., 2013.
[21] P. Tang and T. W. S. Chow, "Wireless sensor-networks conditions monitoring and fault diagnosis using neighborhood hidden conditional random field," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 933–940, Jun. 2016.
[22] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.
[23] W. Li, F. Bassi, D. Dardari, M. Kieffer, and G. Pasolini, "Defective sensor identification for WSNs involving generic local outlier detection tests," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 1, pp. 29–48, Mar. 2016.
[24] A. De Paola, S. Gaglio, G. L. Re, F. Milazzo, and M. Ortolani, "Adaptive distributed outlier detection for WSNs," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 902–913, May 2015.
[25] Y. Tang, B. Zhang, T. Jing, D. Wu, and X. Cheng, "Robust compressive data gathering in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2754–2761, Jun. 2013.
[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[27] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2008.
[28] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.
[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[30] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, 2008.
[31] A. Eriksson, and A. van der Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L_1 norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 771–778.
[32] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
[33] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.
[34] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, May 2011.

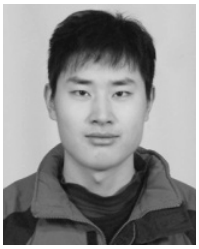
ZHENGYU CHEN received the Ph.D. degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2015. He is currently an Associate Professor with the Jinling Institute of Technology, Nanjing. His current research interests include edge computing, machine learning, and computer communication and networks.

LEI CHEN received the Ph.D. degree in communication and information system from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications. His current research interests include machine learning and computer network.



GUOBING HU received the M.Sc. and Ph.D. degrees in electronic and information engineering from the Nanjing University of Aeronautics and Astronautics, China, in 2006 and 2011, respectively. He is currently an Assistant Professor with the Department of Electronic and Information Engineering, Jinling Institute of Technology, China. His main research interests are radar signal processing and digital signal processing, with a particular focus on noncooperative communications, cognitive radio systems, and multiple-input multiple-output antenna systems.

WENCAI YE received the B.S. degree in computer science and technology from the Nanjing Institute of Technology, Nanjing, China, in 2015. He is currently pursuing the master's degree with the Nanjing University of Posts and Telecommunications, Nanjing. His main research interest is computer communication and networks.



JIN ZHANG received the Ph.D. degree from the College of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016. He is currently a Lecturer with the Department of Electronic and Information Engineering, Jinling Institute of Technology, China. His current research interests include the optimal design of strongly coupled magnetic resonant systems and the parity-time-symmetric method for wireless power transfer.

GENG YANG was born in 1961. He is currenlty a Professor and the Ph.D. Supervisor with the School of Computer Science, Nanjing University of Posts and Telecommunications. His current research interests include computer communication and networks, parallel and distributed computing, and information security.