IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data

YANGWEN YU[1], JAMES J.Q. YU[2,1] (Member, IEEE), VICTOR O.K. LI[1] (Fellow, IEEE), AND JACQUELINE C.K. LAM[1] (Member, IEEE).

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong, China
[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

Corresponding author: James J.Q. Yu (e-mail: yujq3@sustech.edu.cn).

This research is supported in part by the Theme-based Research Scheme of the Research Grants Council of Hong Kong SAR Government, under Grant No. T41-709/17-N.

ABSTRACT With increasing public demands for timely and accurate air pollution reporting, more air quality monitoring stations have been deployed by the governments in urban metropolises to increase the coverage of urban air pollution monitoring. However, due to systematic or accidental failures, some air pollution measurements obtained from these stations are found to have missing values, which will adversely affect the accuracy of any follow-up air pollution analyses and the quality of environmental decision-makings. In this study, the mathematical property of air quality measurements is investigated to recover the missing air pollution values. A new algorithm, which matches meteorology data with air pollution data from different locations, to reconstruct the data matrix and recover missing entries, is proposed. Next, a Low Rank Matrix Completion problem is used to reconstruct the missing values, by transforming the data recovery problem to a sub-gradient primal-dual problem, based on the duality theory, with Singular Value Thresholding (SVT) employed to develop sub-optimal solutions. Next, an Interpolation-SVT (ISVT) approach is adopted to handle the sparsity of observed measurements. Comprehensive case studies are conducted to evaluate the performance of the proposed methods. The simulation results have demonstrated that the proposed SVT and ISVT methods can effectively recover the missing air pollution data and outperform existing interpolation methods and data imputation techniques. The proposed study can improve air pollution estimation and prediction whenever the low-rank data types that are used as proxies for air pollution estimation contain a lot of missing values and require data recovery.

INDEX TERMS Missing data recovery, Interpolation, Low Rank Matrix Completion, Singular Value Thresholding, Air pollution control policy-making.

I. INTRODUCTION

OVER the last few decades, air pollution has presented an increasing health and environmental challenge to many people especially those living in the most populated cities in the world [1]–[3]. Air pollutants, such as $PM_{2.5}$ and PM_{10}, have been responsible for many negative health consequences, such as asthma, Chronic Obstructive Pulmonary Disease (COPD) or cancers [4], [5].

Over the years, governments from all over the world have increasingly tightened up their air pollution control regulations with the hope of reducing air pollution and improve the health of citizens. In China, concerns over continuous deterioration in air qualities throughout the country due to rapid industrialization and economic growth had eventually led to the introduction of a series of highly stringent air pollution control regulations and policies by the central government, starting from the 2000s. The Air Pollution Control Law was updated in 2015/16, explicitly linking public health with air pollution nationally. The updated environmental law has provided a crucial ground for the central and municipal governments to exercise strict controls over emissions generated from coal-fired power plants, and industry and vehicles to cut air pollution at both the municipal and the national level.

Due to the devastating health effects of air pollution, there is an urgent need for timely and accurate disclosure

of ambient air pollution information provided by government monitoring stations. In response to this, an increasing number of air quality monitoring stations have been deployed by the governments around the world in recent years [6]. The quality of air pollution data measured and collected from these stations will directly affect the type of air quality information the publics received at the end, as well as the quality of environmental decision-makings. By ensuring the delivery of good quality air pollution data to the relevant stakeholders, such as the governments or the publics, one can devise relevant measures to reduce the level of air pollution and provide timely health alerts, which are especially crucial for the vulnerable groups such as the asthmatics [7]. Much research has been done to investigate the environmental and health impacts of air pollution [8]–[10].

Nonetheless, such research relies heavily on the integrity of the sampled data obtained from the monitoring stations, and missing air quality data remains an unresolved challenge [11]. The main contributors of data loss include: facility or communication failures, or cyber-security attacks. Firstly, air quality data are typically sampled by electronic devices installed in the monitoring stations, any operational failures may result in data loss [11]. Secondly, communication infrastructure that connects these monitoring stations and the control centers may suffer continuous packet loss or failure [12], [13]. In such case, any data transmitted may not be successfully delivered. Thirdly, exposure of the entire sampling and communication system to network attacks may create potential cyber-security issues [14]. As citizens are increasingly concerned about their personal exposure to ambient air pollution and subsequent health consequences, deliberate attempts to modify or erase these sensitive measurements may occur via cyber attacks. Different drivers of air quality data loss may create different missing data patterns. When an entire measurement station is affected by a communication failure, the missing data will remain spatially and temporally correlated with the sampled data. If a facility and communication failure can be tackled within a short period of time, the missing data are likely to become temporally-correlated. Meanwhile, if any monitoring station suffers from frequent cyber- security attacks or facility failures, missing data may appear to be randomly distributed. In practice, no specific causes will lead to purely spatially-correlated data loss. Furthermore, no data loss satisfying the purely spatially-correlated data loss pattern can be found in the air quality dataset we collected. As such, this study will only focus on the recovery of the first three common types of data loss, namely, (1) the randomly-missing data, (2) the temporally-correlated missing data, and (3) the temporally- and spatially-correlated missing data.

While the missing data in air quality measurements can greatly affect the operation of air quality-related public information services [15], [16], no solution has yet been currently available to address this pressing issue. Directly removing the missing data entries or replacing them with a zero value or with historical data, may affect data distribution and generate biased results, and affect subsequent data analyses. Hence, any simple measures taken to tackle missing air pollution data problem may be undesirable [6].

To bridge this research gap, our previous work [17] first identified the importance of the low rank property of air quality data. We employed Low Rank Matrix Completion (LRMC) to address the missing data challenge for the air quality data. Specifically, we proposed a Singular Value Thresholding (SVT)-based method to tackle the problem. While the previous results demonstrated better performance than the baseline interpolation-based algorithms, the case studies were non-exhaustive. In addition, the previous method was notably challenged by data sparsity, which is commonly found in any real-world air quality samples collected [16]. Under such cases, interpolation-based algorithms can outperform the proposed SVT-based method. By using Interpolation to pre-populate the missing entries into the observation set of SVT, the new Interpolation-SVT (ISVT) algorithm can overcome data sparsity and reconstruct the matrix more effectively. In this new study, the missing rate in [17] will be further re-adjusted. This rate is calculated based on all measurements of the ground truth data [17], instead of the available measurements, as shown in Section IV.

To overcome the matrix reconstruction deficiency of SVT and Interpolation, we propose a new Interpolation SVT-based air quality data recovery algorithm to address the missing data problem. Using Interpolation, we first pre-estimate the missing entries as the additional observations. The additional information is later adopted in SVT to give more accurate estimation.

This research carries the following novelties and significance:

1) We formulate the missing Air Quality Data Recovery (AQDR) problem as an optimization problem, and transform it into a tractable form using interpolation and matrix completion methods.
2) We propose a new strategy to match meteorology data with air pollution data and design an effective algorithm to construct the data matrix and recover missing entries considering the heterogeneous locations of meteorology and air quality data sources.
3) We devise a new method based on SVT and Interpolation algorithms for air quality missing data recovery, which can achieve satisfactory performance on sparsely observed data.
4) We perform a series of comprehensive simulations to assess the performance of the proposed method, and compare our results with existing data imputation and interpolation techniques.

Compared to our previous work re. missing data recovery [17], the current work presents the additional novelties (refer to Point (2) and Point (3) above). Besides, this proposed method can be extended to recover other missing data problems that exhibit similar low rank properties, such as traffic data or meteorology data.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

Recent works aimed at predicting the air quality by using proxy data to fill in the missing information at locations not covered by monitoring stations [6], [18]. However, this work focuses on recovering the missing data directly measured from the stations. There is a significant difference between these two challenges. Furthermore, the recovered air quality data can serve as the input information of the models proposed in [6], [18]. To the best of our knowledge, we are the first team to identify the low-rank property of air pollution data and address the missing data problem in air quality measurements.

The rest of this paper is organized as follows. Section II introduces the formulation of the air quality missing data problem and related work covering missing data recovery with applications on other fields of study. Section III analyses the formulated problem, then elaborates on the details. Section IV investigates the performance of the proposed method. Finally, the conclusion is drawn in Section V, with suggestions on future research.

II. AIR QUALITY DATA RECOVERY PROBLEM

The objective of this work is to recover the missing air quality data, using other observed measurements. This problem can be formulated as an Air Quality Data Recovery (AQDR) problem. We first introduce the AQDR problem. Then we introduce related work that was adopted to recover missing information in other research areas. The definitions and notations generally follow the previous work in [17].

A. NOTATION

In the AQDR problem, we recover the missing entries in a dataset which contains n_{21} categories of air quality measurements sampled from n_{22} monitoring stations in n_1 sampling time intervals. We use $\boldsymbol{M}_{\mathrm{GT}} \in \mathbb{R}^{n_1 \times n_2}$ to denote the ground truth matrix of this dataset, where $n_{21} \times n_{22} = n_2$. Nevertheless, the control center can only obtain a partially observed data matrix, denoted by $\boldsymbol{M}_{\mathrm{OB}}$, instead of $\boldsymbol{M}_{\mathrm{GT}}$ due to data loss:

$$\boldsymbol{M}_{\mathrm{OB}}(i,j) = \begin{cases} \boldsymbol{M}_{\mathrm{GT}}(i,j) & (i,j) \in \boldsymbol{\Omega} \\ \text{null} & \text{otherwise} \end{cases}, \tag{1}$$

where $\boldsymbol{M}_{\mathrm{OB}}(i,j)$ and $\boldsymbol{M}_{\mathrm{GT}}(i,j)$ represent the air quality measurement in matrixes $\boldsymbol{M}_{\mathrm{OB}}$ and $\boldsymbol{M}_{\mathrm{GT}}$, respectively. Ω is the set of matrix indexes (i,j) of the observed entries. For simplicity, we use the following expression to represent the relationship defined in (1):

$$\mathcal{P}_{\Omega}(\boldsymbol{M}_{\mathrm{GT}}) = \mathcal{P}_{\Omega}(\boldsymbol{M}_{\mathrm{OB}}), \tag{2}$$

where $\mathcal{P}_{\Omega}(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is a projector of matrixes that maps indexes in $\boldsymbol{\Omega}$. In addition, we use $\boldsymbol{I}$ to denote the indexes of missing entries in $\boldsymbol{M}_{\mathrm{OB}}$:

$$\boldsymbol{I} = \{(i,j) \notin \boldsymbol{\Omega} | i,j \in \mathbb{Z}^{+}, i \leq n_1, j \leq n_2\}. \tag{3}$$

B. PROBLEM FORMULATION

The objective of the AQDR problem is to estimate the missing data entires in matrix $\boldsymbol{M}_{\mathrm{OB}}$ with the observed data entries. With the notations defined above, AQDR can be formulated as follows:

$$\text{minimize} \ \| \boldsymbol{M} - \boldsymbol{M}_{\mathrm{GT}} \|_2^2 \tag{4a}$$
$$\text{subject to} \ \mathcal{P}_{\Omega}(\boldsymbol{M}) = \mathcal{P}_{\Omega}(\boldsymbol{M}_{\mathrm{GT}}) \tag{4b}$$

where $\boldsymbol{M}$ is the estimated matrix with the recovered entries. In this optimization problem, $\boldsymbol{M}$ should keep all the observations while minimizing the difference from the ground truth $\boldsymbol{M}_{\mathrm{GT}}$ on missing data entries.

However, the ground truth matrix in practice is only partially known in the form of $\boldsymbol{M}_{\mathrm{OB}}$ (or $\mathcal{P}_{\Omega}(\boldsymbol{M}_{\mathrm{GT}})$). When solving problem (4), the lack of $\boldsymbol{M}_{\mathrm{GT}}$ makes the problem intractable, since the only constraint (4b) cannot guarantee a unique solution. In order to get accurate estimations of the missing air quality data, the problem needs to be addressed using other techniques besides this intuitive optimization approach.

C. RELATED WORK

While the missing air quality data can greatly affect the operation of air quality-related public information services [15], [16], no solution has yet been currently identified to address this pressing issue. Some work based on air quality data tends to ignore this missing data problem. For example, the research in [19] simply removed the missing data directly when using deep learning techniques to predict urban air quality. In [6], [18], the missing values were filled in with randomly chosen ones. In some other air quality related previous studies [7], [20], it was common to tackle the missing data by Historical Data Imputation (HDI) or zero imputation, which replaces the lost entries with historical data or zeros. Such methods make the missing data entries either meaningless or inaccurate, which may potentially undermine system performance. More effective methods to solve the AQDR problem are necessary.

1) Interpolation

Interpolation has been employed in a wide range of engineering and science research topics, see [21]–[23] for some examples. As a classical tool for recovering the missing data, interpolation uses a combination of temporally-correlated observations to reconstruct the values of lost data. In real-world applications, a few air quality-related research efforts also attempted to adopt interpolation to address the missing data problem [11], [24]. Two interpolation methods are adopted, between which the Univariate Nearest Neighbor Interpolation (NIN) is arguably the simplest interpolation scheme. This method utilizes the values of the endpoints of the missing gaps to estimate the missing values. However, Linear Interpolation (LIN) adopts the linear fitted line between the endpoints to recover the values of entries in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

missing gap. Both interpolation methods will be investigated in this work.

2) Matrix Completion

Matrix Completion offers an alternative approach to recover missing data [25]. This method focuses on estimating missing entries in the matrix by exploiting the statistical structure of the data. Many methods have demonstrated their efficacies in recovering the ground truth matrix based on Matrix Completion by adopting this characteristics, e.g., total nonlocal models [26], [27], variation-based models [28], [29], sparsity-based models [30], [31], and low-rank models [32], [33].

The low-rank property, which is widely used in practice [32], [34], [35], [36], indicates that the underlying observed matrixes can be represented by linear combinations of a small number of base vectors. Using this property, many low-rank-based matrix completion methods have been proposed. For instance, convex optimization can be employed to estimate the missing entries by minimizing the nuclear norm, e.g., Singular Value Thresholding (SVT) [37] and SoftImpute [38]. Alternatively, matrix factorization with non-convex optimization is also an effective approach. [39]. Considering these recent developments in Low Rank Matrix Completion (LRMC), the algorithm has been widely applied to solve the missing data problem in various engineering research fields, such as in social network [40], power system [41], and remote sensing [42].

3) Other Methods

Apart from interpolation methods, there exist some other approaches for data imputation in other research areas.

For instance, K-Nearest Neighbour (KNN) is an algorithm that is useful for substituting a missing value with its closest k neighbors in a multi-dimensional space [43]. A drawback of this algorithm is the time-consuming nearest neighborhood data point search.

Miss Random Forest (MissForest) imputation is another machine learning based technique that is an extension of classification and regression trees [44]. However, this algorithm aims to predict individual missing data rather than consider the distribution of the dataset. Therefore, this method is not good at handling continuous missing data and the imputed values may lead to biased parameter estimates in statistical models [45].

Expectation maximization (EM) algorithm is an iterative algorithm that finds the maximum log-likelihood of parameters when there are missing values [46]. A drawback of this approach is that it assumes the data loss are uniformly random [47].

Multivariate Imputation by Chained Equations (MICE) is a common method of generating imputed values by drawing from estimated conditional distributions of each variable given all the others [48], [49]. Similar to EM, this approach also requires the missing data positions follow a uniform distribution, which can be unrealistic for practical implementation.

TABLE 1: DATA COLLECTED

Domain	Category	Data Source
Air pollutant	$PM_{2.5}$ ($\mu g/m^3$)	HK government [50], [51]
	PM_{10} ($\mu g/m^3$)	
	O_3 ($\mu g/m^3$)	
	NO_2 ($\mu g/m^3$)	
	SO_2 ($\mu g/m^3$)	
Meteorology	Temperature (°C)	
	Pressure (Pa)	

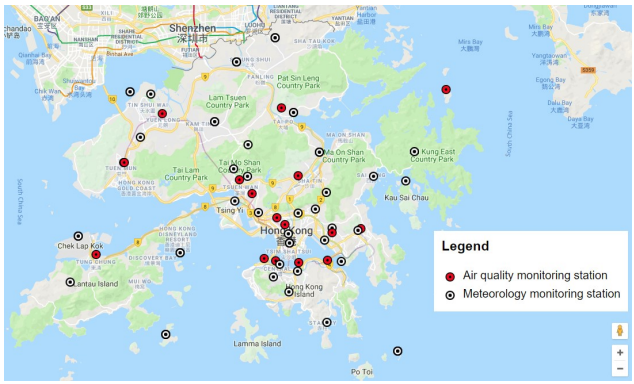


FIGURE 1: Locations of 16 air pollution monitoring stations and 36 meteorology monitoring stations in Hong Kong.

III. DATA RECOVERY MODEL

We firstly introduce a real dataset which suffers from the AQDR problem, and propose a strategy to align the meteorology data with the air pollution data for recovering the missing entries. Then the commonly used methods summarized in Section II will be introduced to address the AQDR problem. By combining the interpolation and the matrix completion approaches, we propose a novel missing data recovery model.

A. THE AIR QUALITY DATASET

The data we used consists of seven categories of air quality related data. As listed in Table 1, all these data are collected from the Hong Kong Environmental Protection Department and Hong Kong Observatory [50], [51]. The first five categories represent the concentrations of five different air pollutants, and the last two are the values of the temperature and pressure, respectively.

Different from the data in [17], these data are sampled synchronously from different official stations distributed across the city of Hong Kong. All five categories of air pollutants data are monitored by 16 air pollution monitoring stations. On the other hand, temperatures are sampled by 36 meteorology monitoring stations, eleven of which also record air pressure values. Fig. 1 shows the locations of both air quality and meteorology monitoring stations in Hong Kong.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

TABLE 2: MISSING VALUES IN THE TEST SET

Source	Number of missing data	Total percentage
$PM_{2.5}/PM_{10}/O_3$ /NO_2/SO_2	410	1.92%
Temperature	45	0.21%
Pressure	96	0.45%
Total	551	2.58%

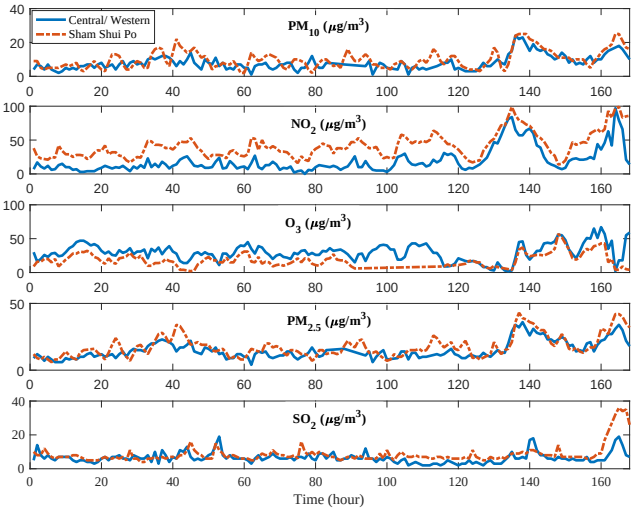


FIGURE 2: Measurements of five air pollutants obtained from the air pollution monitoring stations in Central Western and Sham Shui Po, Hong Kong.

The ground truth dataset contains data collected from 01-Jul-2016 to 07-Jul-2016. All measurements are sampled once per hour. Consequently, 21336 data records are sampled if there is no data loss, i.e., $7 \times 24 \times (16 \times 5 + 36 + 11) = 21336$. However, some data in this dataset were missing due to various events analyzed in Section I. A summary of the missing data is presented in Table 2.

B. MAPPING METEOROLOGY DATA WITH AIR POLLUTION DATA

As depicted in Fig. 2, different measurements of air pollutants sampled by various stations have similar trends within the selected time interval. This observation supports the theoretical analysis presented by the micro-scale dispersion model [52], which indicates significant temporal-spatial (T-S) dependency on the concentrations of air pollutants. In addition, air pollutants can be drastically influenced by urban meteorologies (see [17] and [6] for examples).

Based on the relationship between different categories of air quality related data, the missing data in the dataset can be recovered by utilizing the remaining observed measurements. For instance, in the previous work [17], the observed meteorology data and air pollution data in Beijing can be directly adopted to recover the lost data entries. However, in Hong Kong, the air pollution monitoring stations and the meteorology monitoring stations are not co-located (see Fig. 1). There are different numbers of air pollution and meteorology monitoring stations. In addition, the distance between these stations influence the correlations of different measurements [52]. With the increase of the distance, the correlation decreases. The structural property of the observed data matrix is thus impacted, and missing data recovery is also adversely influenced. In order to improve the recovery accuracy, we propose a data processing algorithm to construct a new observed data matrix $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ from the raw air quality related measurements.

The pseudo-code for constructing of $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ is summarized in Algorithm 1. For each measurement sampled from the p-th meteorology monitoring station, C_p is defined as a correlation coefficient of the meteorology monitoring station with respect to all the other air pollution monitoring stations, which is expressed as follows:

$$C_p = \frac{1}{\sum\limits_q d_{pq}^2}, \tag{5}$$

where d_{pq} denotes the distance between the p-th meteorology monitoring station and the q-th air pollution monitoring station. Similar to the definition of "gates" in other data processing application, e.g., [53], C_p can be regarded as a regulator of the values of the data from the p-th meteorology monitoring station that need to be filled into $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$. Through this regulator, $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ can strengthen the linear correlation of the data by eliminating the negative impact of the distance of stations.

With C_p , $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ can be constructed as follows. Firstly, we can obtain the observed matrix $\boldsymbol{M}_{\mathrm{OB}}$ from the ground truth matrix $\boldsymbol{M}_{\mathrm{GT}}$ of the real measurements dataset. Then, for each $\boldsymbol{M}_{\mathrm{OB}}(i,j)$, which is sampled from the p-th meteorology monitoring station, the corresponding coefficient C_p is employed as a multiplier. Hence, the newly formed matrix $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ can be considered as the new observed matrix for the AQDR problem. Similar to other data processing methods [54], [55], we normalize the data matrix to improve the effectiveness of data processing.

C. INTERPOLATION TO ADDRESS AQDR PROBLEM

By introducing interpolation methods into the AQDR problem, formula (4) can be re-written as

$$\text{minimize} \quad \| \boldsymbol{M}(:,j) - P_{\Theta}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}(:,j)) \|_2^2 \tag{6a}$$

$$\text{subject to} \quad \mathcal{P}_{\Omega}(\boldsymbol{M}) = \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}) \tag{6b}$$

where $j \in \{1, 2, \cdots, n_2\}$. $\boldsymbol{M}(:,j)$, $\widehat{\boldsymbol{M}}_{\mathrm{OB}}(:,j)$ represent the j-th column in $\boldsymbol{M}$ and $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$, respectively. We denote $P_{\Theta}(\cdot)$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

Algorithm 1 Constructing $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$

Input Initial observed matrix $\boldsymbol{M}_{\mathrm{OB}}$, Matrix size (n_1, n_2), Locations of meteorology monitoring stations (x, y) and air pollution monitoring stations (x', y')

for $p = 1$ **to** p_{max}

for $q = 1$ **to** q_{max} **do**

1. $d_{pq} = ((x_p) - (x'_q))^2 + ((y_p) - (y'_q))^2$

end for q

2. $C_p = 1/\sum_{q=1} d_{pq}$

end for p

for $i = 1$ **to** n_1

for $j = 1$ **to** n_2 **do**

3. **if** $\boldsymbol{M}_{\mathrm{OB}}(i,j)$ is sampled from (x_p, y_p)

4. $\widehat{\boldsymbol{M}}_{\mathrm{OB}}(i,j) = C_p \cdot \boldsymbol{M}_{\mathrm{OB}}(i,j)$

5. **else**

6. $\widehat{\boldsymbol{M}}_{\mathrm{OB}}(i,j) = \boldsymbol{M}_{\mathrm{OB}}(i,j)$

7. **end if**

end for j

end for i

Output $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$

as an interpolation-based projection from $\widehat{\boldsymbol{M}}_{\mathrm{OB}}(:,j)$ to $\boldsymbol{M}(:,j)$.

$$P_{\Theta}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}(:,j)) = \begin{cases} \widehat{\boldsymbol{M}}_{\mathrm{OB}}(i,j) & (i,j) \in \Omega \\ \mathrm{Intp}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}(:,j)) & \text{otherwise} \end{cases}, \quad (7)$$

where $\mathrm{Intp}(\cdot)$ represents the rule of the specific interpolation approach. With this rule, the missing entries in the j-th column of $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ can be replaced by the interpolation of observed entries. Based on the different formulations of the Interpolation, $\mathrm{Intp}(\cdot)$ represents different rules. For instance, in NIN, the values of $\mathrm{Intp}(\cdot)$ are the endpoints of the missing gap. Meanwhile, in LIN, $\mathrm{Intp}(\cdot)$ corresponds to the fitted line of the missing gap as introduced in Section II-C. Based on (7), this projector helps to address the AQDR problem.

D. MATRIX COMPLETION FOR TACKLING AQDR PROBLEM

By exploring the structural property of the air quality data, a matrix completion method can be used to solve the AQDR problem. In the past decade, research has been conducted on transforming the matrix completion problem into a convex optimization problem by introducing an extra regularization [56]. For the AQDR problem, in order to obtain a unique and robust solution, equation (4) is reformulated as follows:

$$\text{minimize } R(\boldsymbol{M}) \quad (8a)$$

$$\text{subject to } \mathcal{P}_{\Omega}(\boldsymbol{M}) = \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}) \quad (8b)$$

The regularization term $R(\boldsymbol{M})$ is strategically selected to reflect the structural property of the matrix. With objective function (8a), the optimal solution $\boldsymbol{M}^*$ can accurately resemble the ground truth matrix.

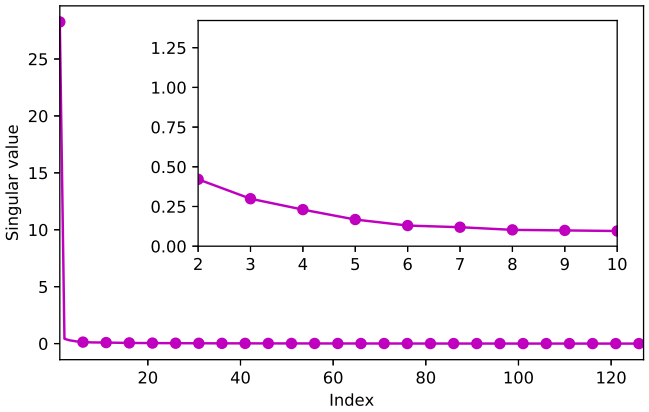


FIGURE 3: Distribution of singular values of $\boldsymbol{M}_{\mathrm{GT}}$.

TABLE 3: THE 10 LARGEST SINGULAR VALUE OF $\boldsymbol{M}_{\mathrm{GT}}$

	σ_i	
Rank	1∼5	6∼10
1	28.2543	0.1299
2	0.4212	0.1195
3	0.2990	0.1029
4	0.2303	0.0993
5	0.1679	0.0956

By LIN, the values of missing entries in $\boldsymbol{M}_{\mathrm{GT}}$ can be filled. Then, utilizing the Singular Value Decomposition (SVD) method [57], the singular values of $\boldsymbol{M}_{\mathrm{GT}}$ can be computed as follows:

$$\boldsymbol{M}_{\mathrm{GT}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad (9)$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrixes, $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix. The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the left and right singular vectors of the matrix, respectively. The non-negative diagonal entries σ_i in $\boldsymbol{\Sigma}$ are the singular values of $\boldsymbol{M}_{\mathrm{GT}}$, which are related to the rank of $\boldsymbol{M}_{\mathrm{GT}}$. The distribution of singular values is shown in Fig. 3. We also list the top-10 largest singular values in Table 3.

As depicted in Fig. 3, the singular values of $\boldsymbol{M}_{\mathrm{GT}}$ diminish quite quickly. This suggests that the matrix can be approximated by a low-rank matrix with high accuracy [58]. Hence, LRMC is suitable for solving the AQDR problem. Equations (4) and (8) can be transformed into a relaxed convex optimization problem [37], [56]:

$$\text{minimize } \tau\|\boldsymbol{M}\|_* + \frac{1}{2}\|\boldsymbol{M}\|_F^2 \quad (10a)$$

$$\text{subject to } \mathcal{P}_{\Omega}(\boldsymbol{M}) = \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}_{\mathrm{OB}}) \quad (10b)$$

where $\|\boldsymbol{M}\|_*$ is the nuclear norm, which is the sum of the singular values of $\boldsymbol{M}$. Additionally, $\|\boldsymbol{M}\|_F$ is the Frobenius norm of $\boldsymbol{M}$. In (10), with a relatively large τ, the solution can be considered as a sub-optimal estimation of $\boldsymbol{M}_{\mathrm{GT}}$ [37]. For this convex optimization problem, the numerical solution can be calculated (see Appendix I), and an SVT can be used to solve it in an iterative manner.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

Based on SVD, we can first define the core operator of SVT as follows, which is named singular value shrinkage:

$$D_\tau(\boldsymbol{X}) \equiv \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}_\tau \boldsymbol{V}^T, \tag{11}$$

where $\tilde{\boldsymbol{\Sigma}}_\tau = \text{diag}\{\sigma_i - \tau\}_+$, and operator $\{t\}_+ = \max(0, t)$. Then, we adopt an intermediate matrix $\boldsymbol{X}$ as the Lagrange Multiplier matrix. By the Sub-gradient Primal-Dual method [59], [60], $\boldsymbol{X}$ can be computed in an iterative manner and adopted into the calculation of $\boldsymbol{M}$. We initialize $\boldsymbol{X}^0 = \boldsymbol{0}$, which, in subsequent iterations $k = 1, 2, \cdots, k_{max}$, are updated using the following rules:

$$\boldsymbol{M}^k = D_\tau(\boldsymbol{X}^{k-1}), \tag{12a}$$

$$\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}} - \boldsymbol{M}^k), \tag{12b}$$

where $\boldsymbol{M}$ is approached iteratively using $\boldsymbol{M}^k$. k_{max} is the maximum number of iterations, which is set to a large positive value (2×10^4 in this paper) [56]. This iterative process repeats until a termination criterion is met:

$$\frac{\|\mathcal{P}_\Omega(\boldsymbol{M}^k - \widehat{\boldsymbol{M}}_{\text{OB}})\|_F}{\|\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}})\|_F} \leq \epsilon, \tag{13}$$

where ϵ is the convergence threshold, which is set to a small positive value 10^{-4} [17]. The pseudo-code for SVT is summarized in Algorithm 2.

Algorithm 2 Singular Value Thresholding (SVT) Algorithm

Input Observed set $\boldsymbol{\Omega}$, observed entries $\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}})$, step size δ, tolerance ϵ, parameter τ, and maximum iteration count k_{max}
Set $\boldsymbol{X}^0 = \boldsymbol{0}$
for $k = 1$ **to** k_{max} **do**
1. $\boldsymbol{M}^k = D_\tau(\boldsymbol{X}^{k-1})$
2. **if** $\|\mathcal{P}_\Omega(\boldsymbol{M}^k - \widehat{\boldsymbol{M}}_{\text{OB}})\|_F / \|\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}})\|_F \leq \epsilon$
3. **break**
4. **else**
5. $\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}} - \boldsymbol{M}^k)$
6. **end if**
end for k
Output $\boldsymbol{M}^k$

E. ISVT ALGORITHM

Similar to the results in [41], [61] and [62], in the previous work [17], the performance of SVT is undermined with the increase of missing rate. For LRMC, there is a lower bound for the observed data. In [56], it is demonstrated that all entries in an $n_1 \times n_2$ matrix, whose rank $r \ll n_1, n_2$, can be efficiently reconstructed with only $O(rn\log_2 n)$ entries , where $n = \max(n_1, n_2)$ [35], [61], [62]. According to the singular value distribution of the ground truth matrix depicted in Fig. 3, for the AQDR problem, $\boldsymbol{M}_{\text{GT}}$ can be approximated by a low-rank matrix with $r \approx 10$. For such matrixes, we need to keep at least around 58.2% overall observed data for SVT to perform well. When addressing cases in which the observed data entries are not sufficient, auxiliary methods are necessary to provide more information for SVT. Combining the interpolation and matrix completion methods, we proposed an Interpolation-SVT (ISVT) algorithm.

Given the missing entry indexes $\boldsymbol{I}$ of $\widehat{\boldsymbol{M}}_{\text{OB}}$, we first randomly select a subset $\boldsymbol{I}_1$ of $\boldsymbol{I}$. Let c, c_1 denote the numbers of elements in $\boldsymbol{I}$ and $\boldsymbol{I}_1$, respectively. We define α as the percentage of missing entries that will be recovered by interpolation initially:

$$\boldsymbol{I}_1 \subset \boldsymbol{I}, \quad \boldsymbol{I}_1 \neq \emptyset, \tag{14}$$

$$c_1 = \alpha c, \quad \alpha \in (0, 1), \tag{15}$$

where α is the proportion of c_1 over c. In ISVT, the missing entries in $\boldsymbol{I}_1$ are recovered by the interpolation method in the first step. Then these recovered entries are added into Ω, which is subsequently processed by SVT. When handling different missing rate cases, α may have different optimal values to achieve the best performance. The selection of α will be investigated in Section IV-B. The pseudo-code of the proposed ISVT algorithm is as shown in Algorithm 3. The convergence analysis of the proposed algorithm is presented in Appendix II.

Algorithm 3 ISVT Algorithm

Input Observed set $\boldsymbol{\Omega}$, observed entries $\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}})$, missing entries indexes set $\boldsymbol{I}$, missing entries counts c, percentage α , step size δ, tolerance ϵ, parameter τ, and maximum iteration count k_{max}
1, **Select** $\boldsymbol{I}_1$ from $\boldsymbol{I}$ based on c and α
2. **Initialization** Interpolation for entries located in $\boldsymbol{I}_1$
3. **Combine** $\boldsymbol{I}_1$ and $\boldsymbol{\Omega}$ **into** $\tilde{\boldsymbol{\Omega}}$
4. **Add** entries located in $\boldsymbol{I}_1$ into $\widehat{\boldsymbol{M}}_{\text{OB}}$, we have $P_{\tilde{\boldsymbol{\Omega}}}(\widehat{\boldsymbol{M}}_{\text{OB}})$
5. **Set** $\boldsymbol{X}^0 = \boldsymbol{0}$
6.1 **for** $k = 1$ **to** k_{max} **do**
6.2 $\boldsymbol{M}^k = D_\tau(\boldsymbol{X}^{k-1})$
6.3 **if** $\|\mathcal{P}_\Omega(\boldsymbol{M}^k - \widehat{\boldsymbol{M}}_{\text{OB}})\|_F / \|\mathcal{P}_\Omega(\widehat{\boldsymbol{M}}_{\text{OB}})\|_F \leq \epsilon$
6.4 **break**
6.5 **else**
6.6 $\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta P_{\tilde{\boldsymbol{\Omega}}}(\widehat{\boldsymbol{M}}_{\text{OB}} - \boldsymbol{M}^k)$
6.7 **end if**
6.8 **end for** k

7. **Output** $\boldsymbol{M}^k$

IV. EXPERIMENT

In order to demonstrate the performance of the proposed algorithms in recovering missing air quality data, a series of simulations are conducted. In these experiments, we use the normalized $\boldsymbol{M}_{\text{GT}}$ as described in Section III as the ground truth matrix. Besides the original 551 missing data, some other air pollution entries are erased in different cases, to mimic missing data. We use p to represent the ratio of the missing data to overall measurements.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

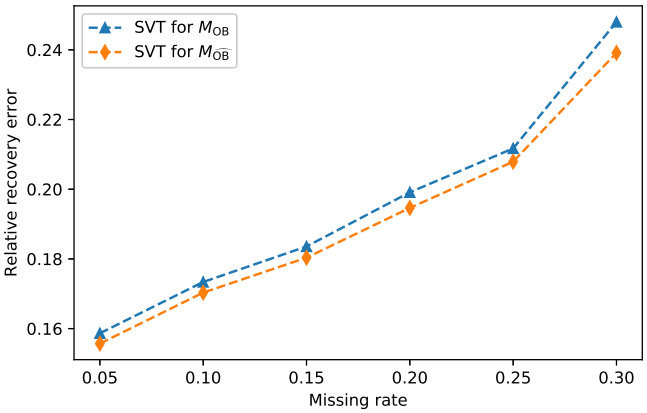


FIGURE 4: Comparision of $\boldsymbol{M}_{\mathrm{OB}}$ and $\boldsymbol{M}_{\widehat{\mathrm{OB}}}$

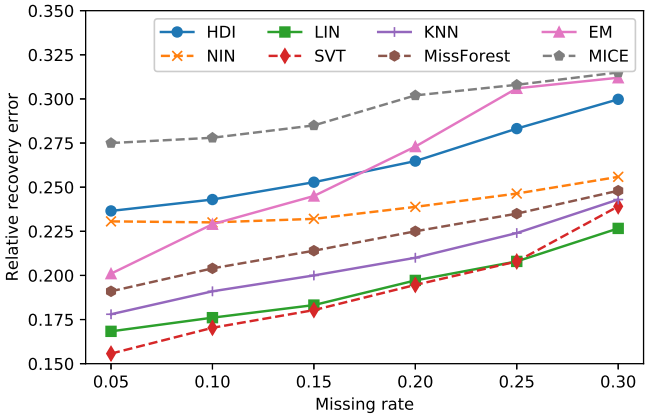


FIGURE 5: Relative recovery errors of four methods for randomly missing air pollution data entries.

The recovery performance of the AQDR problem is measured by the relative recovery error, which has been adopted in previous literatures, e.g., [17] and [41]:

$$Re = \frac{\|\boldsymbol{E} - \boldsymbol{E}_{\mathrm{rec}}\|_F}{\|\boldsymbol{E}\|_F}, \tag{16}$$

where $\boldsymbol{E}$ represents the real values of the erasures in $\boldsymbol{M}_{\mathrm{GT}}$, and $\boldsymbol{E}_{\mathrm{rec}}$ represents the recovered values of these data entries.

The sensitivity analysis of SVT has been illustrated in the previous work [17]. For simplicity, we empirically set these parameters according to their overall performance in different missing data scenarios [17]. The values of the SVT parameters are as follows:

$$\tau = 1.2\sqrt{n_1 n_2}, \tag{17}$$

$$\delta = 1.5\frac{n_1 n_2}{|\Omega|}, \tag{18}$$

In addition, the convergence threshold ϵ in SVT is set to 10^{-4} [17].

A. PERFORMANCE OF PROPOSED $\widehat{M}_{\mathrm{OB}}$

This subsection aims to evaluate $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$, which is obtained by the proposed construction method. In order to investigate the performance of the constructed matrix, the SVT algorithm is employed to recover the missing data. With different values of $p \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, we randomly remove some entries of air pollution data in $\boldsymbol{M}_{\mathrm{GT}}$ to obtain the initial observed matrix $\boldsymbol{M}_{\mathrm{OB}}$. By Algorithm 1, $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ can be developed from $\boldsymbol{M}_{\mathrm{OB}}$. The relative recovery errors of the erasures under different p values are presented in Fig. 4. The performance of the initial observed matrix $\boldsymbol{M}_{\mathrm{OB}}$ is also displayed in this figure as a baseline performance.

As shown in Fig. 4, it can be observed that the proposed $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$, which maps meteorology data into air pollution data, outperforms the initial $\boldsymbol{M}_{\mathrm{OB}}$. The performance improvement of $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ over $\boldsymbol{M}_{\mathrm{OB}}$ persists regardless of the missing rate p. It can be concluded that the data processing algorithm can significantly improve the recovery performance of SVT.

B. COMPUTATIONAL COMPLEXITY ANALYSIS

For any matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2} (n_1 \geqslant n_2)$, the computational complexity of SVD is $O(n_1 n_2^2)$ [63]. As the core operator can be regarded as a truncated SVD, the major computational cost of SVT in each iteration is contributed by SVD. Compared to SVT, ISVT injects an interpolation process in front of the subsequent SVT computation. The complexity of interpolation is $O(n_2)$. Therefore, the computation cost of ISVT is $O(n_2 + kn_1 n_2^2)$, where k is the number of iterations. Since $kn_1 n_2^2 \gg n_2$, ISVT spends the longest time on addressing the optimization problem (19) by SVT iteration. From the above discussion, we can conclude that the computational complexity of ISVT is comparable to (in the same order of magnitude) that of SVT.

From the computational complexity analysis, it is obvious that the time consumption of SVT increases dramatically with the size of data. So, in the previous studies, some researchers focus on the acceleration of matrix completion [64], [65]. However, such acceleration is at the expense of data reconstruction accuracy. Since the scale of the data we used is relatively small ($\min(n_1, n_2) < 1000$), the time consumption of traditional SVT is affordable (within several minutes). And this computation time is much smaller than the monitoring time interval (1hr). Thus, we can conclude that both SVT and ISVT can recover the weekly air measurement matrix in time.

C. RECOVERY OF RANDOMLY MISSING DATA

The performance of the proposed SVT and ISVT in recovering randomly missing entries are assessed. In this test, we randomly remove some air pollution data entries in $\boldsymbol{M}_{\mathrm{GT}}$ to construct $\boldsymbol{M}_{\mathrm{OB}}$. Then according to the result of Section IV-A, the better-performing $\widehat{\boldsymbol{M}}_{\mathrm{OB}}$ by Algorithm 1 is adopted for recovering the missing measurements. With different $p \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, all results are averaged based on 50 independent runs, in which the erasures are randomly generated for each run. The simulation results are depicted in Fig. 5 and Fig. 6.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

TABLE 4: RELATIONSHIP BETWEEN MISSING RATE p AND RATIO α

Re	α of ISVT											SVT	LIN
p	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%		
0.05	**0.1557**	0.1558	0.1568	0.1578	0.1577	0.1595	0.1614	0.1625	0.1629	0.1633	0.1641	0.1557	0.1683
0.10	**0.1703**	0.1705	0.1707	0.1711	0.1721	0.1714	0.1723	0.1721	0.1734	0.1736	0.1742	0.1703	0.1761
0.15	0.1803	**0.1801**	0.1804	0.1807	0.1805	0.1803	0.1808	0.1808	0.1806	0.1810	0.1817	0.1803	0.1832
0.20	0.1946	0.1941	0.1935	0.1938	0.1936	0.1933	0.1921	0.1922	**0.1920**	0.1925	0.1935	0.1946	0.1972
0.25	0.2074	0.2070	0.2057	0.2036	0.2030	0.2026	0.2019	**0.2011**	0.2027	0.2047	0.2056	0.2079	0.2079
0.30	0.2358	0.2277	0.2230	**0.2195**	0.2210	0.2207	0.2206	0.2210	0.2216	0.2231	0.2244	0.2391	0.2266
0.40	0.9717	0.8065	0.6385	0.3472	0.2774	**0.2740**	0.2793	0.2865	0.2898	0.2960	0.3000	-	0.3026
0.50	-*	-	-	-	0.6115	0.3927	**0.3791**	0.3898	0.4006	0.4088	0.4172	-	0.4219

* — represent the value is larger than 1.0, it can be considered as an invalid value of relative recovery error.

Fig. 5 shows that, even though the recovery error of each algorithm increases with the missing rates, LIN and SVT consistently surpass other approaches across the entire range of missing rates. When $p \leq 0.25$, SVT significantly outperforms LIN. Meanwhile, when the missing rate is extremely high ($p > 0.25$), LIN performs slightly better. However, it is highly unlikely that the data suffers from such a high missing rate in practice [6]. Therefore, it can be concluded that SVT generally achieves better missing data recovery performance as compared to MICE, HDI, NIN, EM, MissForest, KNN and LIN. The outstanding performance of SVT is attributed to its ability to recover multivariate data. Except for using the observed data, nuclear-norm based optimization can estimate the multivariate missing entries by fully exploring the low-rank property of the entire data distribution. However, the recovery accuracy of other traditional methods can be severely degraded with high-dimensional data [45], [47], [66].

As analyzed in Section III-E, the main cause of the decreased accuracy on high missing rate is the sparsity of the observed data. In such cases, SVT is limited by the number of observations. When the number of observations is below the suggested lower bound [56], namely, approximately 58.2% c.f. Section III-E, SVT has subpar recovery accuracy performance. Considering this factor, with linear interpolation method that has the sub-optimal performance, ISVT may improve the performance by introducing partially recovered missing entries into the observations.

We also assess the performance of ISVT with different α values. For each p, 50 independent runs are evaluated. As shown in Fig. 6, the best-performing ISVT (ISVT based on LIN) outperforms both SVT and LIN. In particular, when $p > 0.25$, the superiority of the best-performing ISVT over SVT is more obvious. The averaged relative recovery errors of ISVT based on LIN are presented in Table 4 with different α and p, where the best performing α values are in bold. Even for extreme cases ($p \geq 0.3$), ISVT can still achieve better performance than other methods. Moreover, with the increase of missing rate p, a relatively large α is shown to be much more effective.

D. RECOVERY OF TEMPORALLY-CORRELATED MISSING DATA

In the real world, due to facility failures and other issues, consecutive data loss is commonly encountered. In these

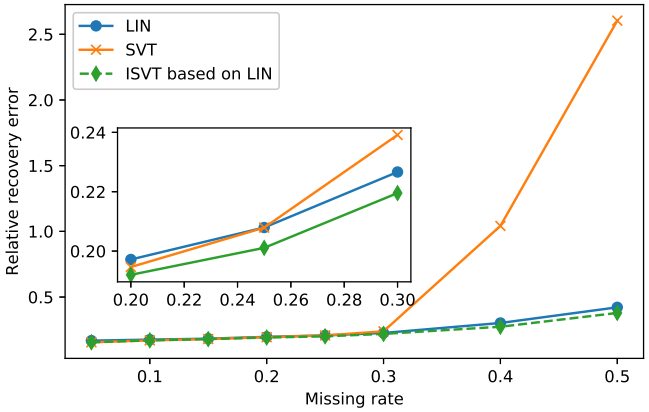


FIGURE 6: Comparison of ISVT, SVT and LIN

TABLE 5: RELATIVE RECOVERY ERRORS FOR 10% TIME CORRELATED RANDOMLY MISSING AIR POLLUTION DATA ENTRIES

	N_t				
Recovery methods	2	4	8	16	32
KNN	0.2231	0.2523	0.2892	0.3083	0.3322
MissForest	0.2252	0.2434	0.2714	0.2882	0.3131
EM	0.3232	0.3533	0.3680	0.3730	0.3844
MICE	0.3043	0.3081	0.3120	0.3172	0.3273
LIN	0.2146	0.2832	0.4840	0.8592	1.7051
SVT	0.1872	**0.1989**	**0.2186**	**0.2310**	**0.2537**
ISVT	**0.1871**	0.2010	0.2214	0.2437	0.2892

cases, the missing data are temporally correlated. In order to investigate the performance of the proposed methods for recovering such incomplete air pollution data, we develop the observed data matrix by removing consecutive samples from $\boldsymbol{M}_{\mathrm{GT}}$. By setting the missing rate p constant, we remove consecutive air pollution data sequences in $\boldsymbol{M}_{\mathrm{GT}}$ randomly. Let $N_t \in \{2, 4, 8, 16, 32\}$ be the number of consecutive entries in one temporal erasures sequence. To focus on the performance of ISVT in handling such cases, the locations of α missing data are included in $\boldsymbol{I}_1$ for LIN initially. All simulations are conducted for 50 times for statistical significance.

The averaged recovery errors for $p = 0.1$ and $p = 0.3$ are presented in Fig. 7. The plots of ISVT represent the recovery performance of the best-performing ISVT based on LIN. Tables 5 and 6 show more detailed comparisons of the recovery accuracy of seven different approaches, namely

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

TABLE 6: RELATIVE RECOVERY ERRORS FOR 30% TIME CORRELATED RANDOMLY MISSING AIR POLLUTION DATA ENTRIES

	N_t				
Recovery methods	2	4	8	16	32
KNN	0.2541	0.2910	0.3253	0.3764	0.3893
MissForest	0.2500	0.2712	**0.3011**	0.3524	0.3772
EM	0.3469	0.3518	0.3565	0.3719	0.3916
MICE	0.3122	0.3210	0.3451	0.3613	0.3800
LIN	0.2627	0.3753	0.6841	1.3261	3.0072
SVT	0.2858	0.3043	0.3299	**0.3464**	**0.3720**
ISVT	**0.2287**	**0.2635**	0.3042	0.3562	0.4357

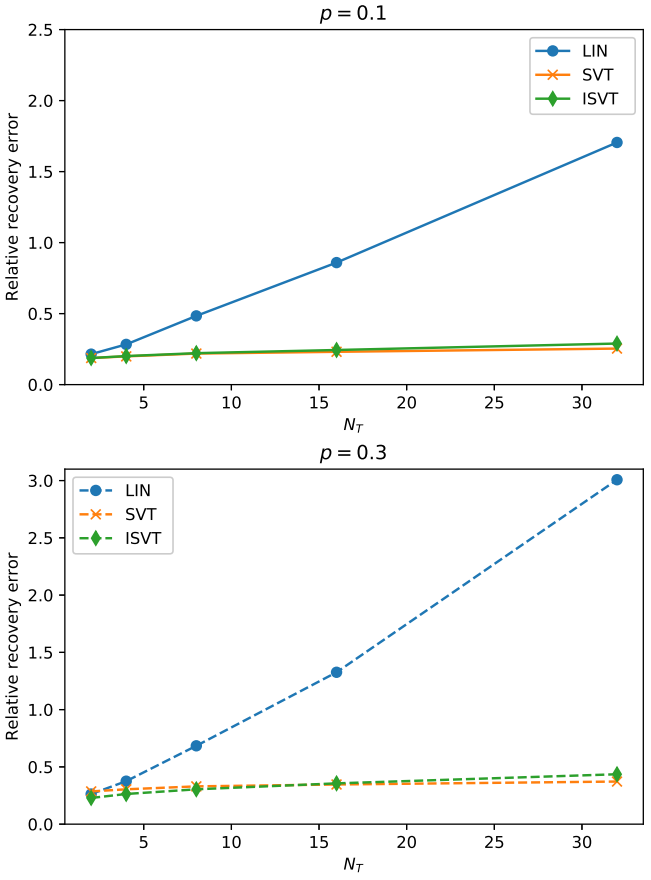


FIGURE 7: Relative recovery errors for 10% and 30% temporally-correlated randomly missing air pollutant data entries.

KNN, MissForest, EM, MICE, LIN, SVT and ISVT. With the increase of p, the performance of all seven methods become worse. At the same time, for a fixed missing rate, the recovery accuracy of the methods are also undermined with the increase of the value of N_t. The deteriorating performance of traditional methods is caused by their strict random missing data assumptions. Also, due to the decrease of observations, the performance of SVT-based methods is limited as discussed in Section III-E.

As depicted in Table 5, although ISVT has a slightly better performance when $p = 0.1$, $N_t = 2$ (see Table 5), SVT can be considered as the best performing method compared to other methods in the other N_t cases. This is due to the bad performance of LIN in the time-related missing data problem. By LIN, inaccurate data entries are used in ISVT as the input observations. This makes ISVT worse than SVT. The difference between the relative recovery error of SVT and ISVT becomes more significant due to the rapidly deteriorated LIN performance.

The impact of LIN on ISVT is more significant for $p = 0.3$. Table 6 shows that, although LIN can obviously improve the performance of ISVT compared to SVT when $N_t \leq 4$, the recovery accuracies are severely constrained by the drawback of LIN with the increase of N_t. Although MissForest outperforms SVT at $N_t = 8$, similar to LIN, MissForest cannot handle continuous miss data cases. SVT is also the best choice in cases with other N_t value.

Therefore, we can conclude that SVT is the best method for handling the temporally-correlated missing data problem. However, for the extremely large missing case, the benefits of ISVT cannot be ignored.

E. RECOVERY OF TEMPORALLY- AND SPATIALLY-CORRELATED MISSING DATA

As introduced in Section I, different types of air pollution monitoring equipment are deployed in the same stations. It is possible that some monitoring devices in one station are broken at the same time. Such missing data may exhibit spatial-correlation characteristics. In this case, the missing data are not only temporally but also spatially correlated. These lost entries in $\boldsymbol{M}_{GT}$ can be considered as "blocks" . By setting p constant, we randomly remove some fixed size data blocks in $\boldsymbol{M}_{\mathrm{GT}}$ to construct $\boldsymbol{M}_{\mathrm{OB}}$. Besides the original missing data, we use symbol $N_{ts} = N_t \times N_s$ to indicate the size of one missing block, where N_s denotes the number of columns that suffer from the spatially correlated data loss.

We set $N_t \in \{2, 4, 8, 16\}$ and $N_s \in \{2, 3\}$. So there are totally eight different sizes of missing blocks. For each N_{ts}, the simulations are conducted 50 times for statistical significance. The averaged recovery errors are presented in Fig. 8, Tables 7 and 8 for $p \in \{0.1, 0.3\}$.

As shown in Fig.8, LIN is less effective in addressing temporally-spatially correlated missing data. In addition, SVT is always better than the best performing ISVT with LIN except for the case that $p = 0.1$, $N_t = 2$. With the increase of N_t, the advantage becomes more obvious. This observation accords with the previous results in Section IV-D.

Furthermore, Tables 7 and 8 give the detailed comparison of different methods. From the tables, it is clear that the difference of the recovery accuracy between SVT and other approaches increases not only with N_t but also N_s. Since SVT recovers the missing data by exploiting the structural property of the whole data matrix, SVT is better than LIN, which only adopts the endpoint values of the missing gap for estimating spatially correlated missing data. Furthermore, as limited by the individual drawbacks described in Section II, other baseline methods are also worse than SVT for recov-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

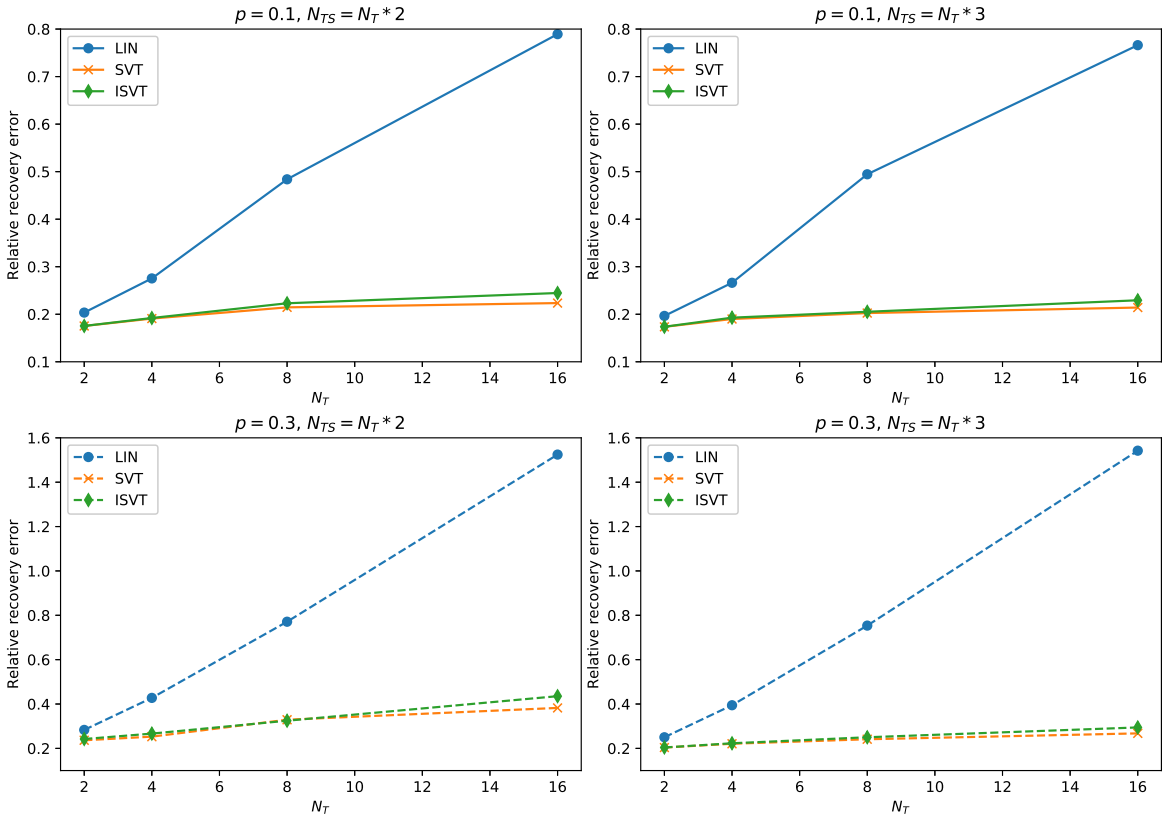


FIGURE 8: Relative recovery errors for 10% and 30% temporally-spatially correlated randomly missing air pollutant data entries.

TABLE 7: RELATIVE RECOVERY ERRORS FOR 10% TEMPORALLY-SPATIALLY CORRELATED RANDOMLY MISSING AIR POLLUTION DATA ENTRIES

	$N_s = 2$				$N_s = 3$			
N_t	2	4	8	16	2	4	8	16
KNN	0.2083	0.2383	0.2800	0.2944	0.2372	0.2534	0.2816	0.2889
MissForest	0.2114	0.2273	0.2587	0.2769	0.2414	0.2487	0.2705	0.2778
EM	0.3421	0.3516	0.3565	0.3731	0.3042	0.3424	0.3729	0.4026
MICE	0.2832	0.2923	0.3091	0.3118	0.2916	0.2977	0.3099	0.3148
LIN	0.2035	0.2754	0.4839	0.7893	0.2022	0.2762	0.4845	0.7761
SVT	0.1753	**0.1909**	**0.2145**	**0.2235**	0.1735	**0.1900**	**0.2065**	**0.2141**
ISVT	**0.1751**	0.1921	0.2232	0.2447	**0.1731**	0.1928	0.2152	0.2294

ering the continuous missing data. Besides, the performance of ISVT is influenced by the inaccurate LIN computation. Hence for temporally-spatially missing data scenarios, SVT has the best performance.

In summary, the previous case studies provide guidelines for selecting the best recovery methods to address different missing data patterns. Specifically, ISVT has the best performance for randomly missing data, and SVT is considered best for temporally- and temporally-spatially correlated missing data. In addition, it can be observed that for all missing data patterns, SVT is more stable. Although the effectiveness of ISVT cannot be ignored, LIN limits the performance of ISVT in continuous data loss cases. As previously analyzed, SVT is insensitive to the distribution of missing data since it can utilize complete information of the matrix, while the traditional interpolation relies on merely the neighbor entries around the missing values, leading to a relatively unstable performance.

V. CONCLUSION

This study examines the problem of missing data recovery, using air pollution data recovery as a case study. The problem has been formulated and two widely used data recovery approaches, namely, the Interpolation approach and the Matrix Completion approach, have been introduced. Given the heterogeneous distribution of monitoring stations for air pollution and meteorology, a new strategy to reconstruct the data matrix to recover the missing air pollution data has been proposed. Next, the low-rank property of a newly constructed data matrix has been introduced. The formulated AQDR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

TABLE 8: RELATIVE RECOVERY ERRORS FOR 30% TEMPORALLY-SPATIALLY CORRELATED RANDOMLY MISSING AIR POLLUTION DATA ENTRIES

	$N_s = 2$				$N_s = 3$			
N_t	2	4	8	16	2	4	8	16
KNN	0.2722	0.2773	0.3497	0.3968	0.2496	0.2638	0.2929	0.3271
MissForest	0.2681	0.2986	0.3280	0.3842	0.2377	0.2735	0.3161	0.3754
EM	0.3044	0.3328	0.4651	0.5646	0.3000	0.3132	0.3814	0.5261
MICE	0.2926	0.3180	0.3597	0.3984	0.2823	0.3066	0.3127	0.3340
LIN	0.2832	0.4276	0.7710	1.5245	0.2501	0.3945	0.7535	1.5421
SVT	**0.2365**	**0.2529**	**0.3247**	**0.3824**	**0.2037**	**0.2209**	**0.2411**	**0.2676**
ISVT	0.2422	0.2666	0.3292	0.4353	0.2041	0.2230	0.2501	0.2944

problem can be transformed into an LRMC problem. The relaxed convex form of the LRMC problem has a numerical solution, with SVT generating an iterative procedure to solve the challenge. In an extreme case, when the observed data is too sparse to ensure the recovery accuracy of SVT, the new ISVT algorithm, which incorporates the values of the missing entries estimated by Interpolation into the observations, is developed.

Simulations have been conducted to test the performances of the proposed SVT and ISVT. The simulation results indicate that SVT and ISVT outperform the traditional interpolation methods and data imputation techniques in most of our test cases. In addition, with the increase in data missing rates, more intermediate data can be estimated by Interpolation to obtain a better performance. Therefore, ISVT has a better performance when recovering randomly missing air quality values. With sparse observations for handling the temporally-correlated or the temporally- and spatially- correlated missing air pollution data, SVT is the best choice.

Future studies will focus on the design of novel missing data recovery approaches based on advanced deep learning methods. Furthermore, our proposed approach may also be extended to other research problems. While the low rank property of air quality data is dependent on the high spatial-temporal (S-T) correlations [17], many other naturally occurring data, such as meteorology data, traffic data, or power system data may also exhibit similar spatial-temporal (S-T) correlations characteristics. It is highly plausible that this newly proposed method, which utilizes the respective low rank properties, can be applied in such contexts for missing data recovery.

.

APPENDIX I. NUMERICAL SOLUTION FOR ADDRESSING LRMC

Based on the modified objective function, the original AQDR problem can be transformed into a convex optimization problem:

$$\text{minimize } \tau\|\boldsymbol{M}\|_* + \frac{1}{2}\|\boldsymbol{M}\|_F^2 \tag{19a}$$

$$\text{subject to } \mathcal{P}_\Omega(\boldsymbol{M}) = \mathcal{P}_\Omega(\boldsymbol{M}_{\text{GT}}) \tag{19b}$$

where the parameter $\tau > 0$ is constant. The corresponding Lagrangian can be constructed as follows :

$$\begin{aligned}\mathcal{L}(\boldsymbol{M}, \boldsymbol{X}) =& \tau\|\boldsymbol{M}\|_* + \frac{1}{2}\|\boldsymbol{M}\|_F^2 \\ &+ \langle \boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{M}_{\text{GT}}) - \mathcal{P}_\Omega(\boldsymbol{M})\rangle,\end{aligned} \tag{20}$$

$\boldsymbol{X}$ corresponds to the Lagrangian, and $\langle A, B\rangle = \text{trace}(A * B)$.

In addition, the objective function (19a) is convex, and the sole constraint (19b) is a linear equality. This makes the optimization problem accord with Slater's conditions [67]. Hence, the strong duality property holds. Using (20), we can solve the modified problem numerically with the sub-gradient method [59], [60]. For the primal problem (19), the corresponding dual problem is constructed as follows:

$$g(\boldsymbol{X}) = \inf_M \mathcal{L}(\boldsymbol{M}, \boldsymbol{X}). \tag{21}$$

By the sub-gradient method, the Lagrange multiplier $\boldsymbol{X}$ can be computed in an iterative manner:

$$\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta h^{(k-1)}, \tag{22}$$

where δ is a positive step size, and $\boldsymbol{X}$ is initialized as $\boldsymbol{X}^0 = \boldsymbol{0}$. $h^{(k-1)}$ is the sub-gradient of the dual problem at $\boldsymbol{X}^{k-1}$, which provides a gradient descent direction for $\boldsymbol{X}$. In subsequent iterations $k = 1, 2, \cdots$, we consider $\boldsymbol{M}^{k^*}$ as optimal for $\mathcal{L}(\boldsymbol{M}^k, \boldsymbol{X}^k)$. Hence, at point $\boldsymbol{X}^{k-1}$, the sub-gradient $h^{(k-1)}$ for $\boldsymbol{X}$ can be calculated as

$$\begin{aligned}h^{(k-1)} &= \mathcal{P}_\Omega(\boldsymbol{M}_{\text{GT}}) - \mathcal{P}_\Omega(\boldsymbol{M}^{(k-1)^*}) \\ &= \mathcal{P}_\Omega(\boldsymbol{M}_{\text{GT}} - \boldsymbol{M}^{(k-1)^*}) \\ &= \mathcal{P}_\Omega(\boldsymbol{M}_{\text{OB}} - \boldsymbol{M}^{(k-1)^*}).\end{aligned} \tag{23}$$

After updating $\boldsymbol{X}$, we also need to update the optimal $\boldsymbol{M}^{k^*}$ inductively:

$$\begin{aligned}\boldsymbol{M}^k &= \boldsymbol{M}^{(k-1)^*} \\ &= \text{argmin} \quad \mathcal{L}(\boldsymbol{M}, \boldsymbol{X}^{k-1}).\end{aligned} \tag{24}$$

Combining (21)–(24), $\boldsymbol{X}$ and $\boldsymbol{M}$ are updated by the following rules:

$$\boldsymbol{M}^k = \text{argmin} \quad \mathcal{L}(\boldsymbol{M}, \boldsymbol{X}^{k-1}), \tag{25a}$$

$$\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta\mathcal{P}_\Omega(\boldsymbol{M}_{\text{OB}} - \boldsymbol{M}^k). \tag{25b}$$

The estimated values of missing entries can be updated from $\boldsymbol{M}^k$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2988684, IEEE Access

APPENDIX II. CONVERGENCE ANALYSIS

We will analyze the convergence for the algorithm. In ISVT, the interpolation operation and SVT cascade. Although the pre-interpolation operation can enhance the input observations of subsequent SVT, the optimization problem (19) is not changed. Therefore, the convergence of ISVT algorithm is based on the convergence of the iterative SVT for our optimization problem.

Let $f(\cdot)$ be the original objective function (19a) and $\partial f(\boldsymbol{M})$ be a subgradient of f at $\boldsymbol{M}$. First, we establish the convexity of the objective $f(\cdot)$ by proving the convexity of both $\|\cdot\|_*$ and $\|\cdot\|_F^2$, respectively.

Lemma 1: The objective function $f(\cdot)$ is convex.

Proof 1 (Proof): Let $f_1(\cdot) = \|\cdot\|_*$ and $f_2(\cdot) = \|\cdot\|_F^2$. The subgradients of f_1 and f_2 at $\boldsymbol{M}$ are $\partial f_1(\boldsymbol{M})$ and $\partial f_2(\boldsymbol{M})$, respectively. Then

$$\partial f_1(\boldsymbol{M}) = \partial\|\boldsymbol{M}\|_*, \tag{26}$$

$$\partial f_2(\boldsymbol{M}) = \partial\|\boldsymbol{M}\|_F^2 = 2\boldsymbol{M}, \tag{27}$$

Thus, for $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n_1\times n_2}$, we get

$$\langle\partial f_2(\boldsymbol{M}_1) - \partial f_2(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle = 2\|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_F^2 \geqslant 0, \tag{28}$$

which proves the convexity of f_2.

Besides, From Lemma 4.1 in [56], we obtained that $\|\partial f_1(\boldsymbol{M})\|_2 \leqslant 1$ and $\langle\partial f_1(\boldsymbol{M}), \boldsymbol{M}\rangle = \|\boldsymbol{M}\|_*$ at any $\boldsymbol{M}$. For any $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n_1\times n_2}$, this gives

$$|\langle\partial f_1(\boldsymbol{M}_1), \boldsymbol{M}_2\rangle| \leqslant \|\partial f_1(\boldsymbol{M}_1)\|_2\|\boldsymbol{M}_2\| \leqslant \|\boldsymbol{M}_2\|_*, \tag{29}$$

Therefore,

$$\begin{aligned}\langle\partial f_1(\boldsymbol{M}_1)-\partial f_1(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle &= \|\boldsymbol{M}_1\|_* + \|\boldsymbol{M}_2\|_* \\ &- \langle\partial f_1(\boldsymbol{M}_1), \boldsymbol{M}_2\rangle - \langle\partial f_1(\boldsymbol{M}_2), \boldsymbol{M}_1\rangle \geqslant 0,\end{aligned} \tag{30}$$

which proves the convexity of f_1.

Since f ($f = \tau f_1 + \frac{1}{2}f_2$) is a linear combination of the convex functions f_1 and f_2, f is convex.

This lemma is essential for the convergence of this algorithm. According to Lemma 1, we show the boundedness of $\langle\partial f(\boldsymbol{M}_1) - \partial f(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle$.

Lemma 2: Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n_1\times n_2}$. Then

$$\langle\partial f(\boldsymbol{M}_1) - \partial f(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle \geqslant \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_F^2 \tag{31}$$

Proof 2: Since $f = \tau f_1 + \frac{1}{2}f_2$, the left-hand term can be rewritten as

$$\begin{aligned}\langle\partial f(\boldsymbol{M}_1) - \partial f(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle &= \tau\langle\partial f_1(\boldsymbol{M}_1) \\ -\partial f_1(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle &+ \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_F^2\end{aligned} \tag{32}$$

According to inequality (30), $\langle\partial f_1(\boldsymbol{M}_1) - \partial f_1(\boldsymbol{M}_2), \boldsymbol{M}_1 - \boldsymbol{M}_2\rangle \geqslant 0$. Thus, Lemma 2 holds.

With these lemmas, we now prove the convergence of the algorithm.

Theorem 1: The updated $\boldsymbol{M}^k$ in (25) can converge to the unique solution of our optimization problem (19), if the step size δ satisfies $0 < \delta < 2$.

Proof 3 (Proof): For some iteration k, $\boldsymbol{M}^k$ is the minimizer of $\mathcal{L}(\boldsymbol{M}, \boldsymbol{X}^{k-1})$ as given in (24). Therefore,

$$\nabla\mathcal{L}(\boldsymbol{M}^k, \boldsymbol{X}^{k-1}) = \partial f(\boldsymbol{M}^k) - \mathcal{P}_\Omega(\boldsymbol{X}^{k-1}) = 0 \tag{33}$$

Let $(\boldsymbol{M}^*, \boldsymbol{X}^*)$ be the primal-dual optimal for the problem (19). Based on the Karush–Kuhn–Tucker (KKT) conditions [68], we have

$$\nabla\mathcal{L}(\boldsymbol{M}^*, \boldsymbol{X}^*) = \partial f(\boldsymbol{M}^*) - \mathcal{P}_\Omega(\boldsymbol{X}^*) = 0 \tag{34}$$

Based on Lemma 2, we have

$$\begin{aligned}&\langle\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*), \boldsymbol{M}^k - \boldsymbol{M}^*\rangle = \\ &\quad\langle\partial f(\boldsymbol{M}^k) - \partial f(\boldsymbol{M}^*), \boldsymbol{M}^k - \boldsymbol{M}^*\rangle \geqslant \|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2\end{aligned} \tag{35}$$

Since $\mathcal{P}_\Omega(\boldsymbol{M}^*) = \mathcal{P}_\Omega(\boldsymbol{M}_{OB})$ and $\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta\mathcal{P}_\Omega(\boldsymbol{M}_{\mathrm{OB}} - \boldsymbol{M}^k)$,

$$\begin{aligned}&\|\mathcal{P}_\Omega(\boldsymbol{X}^k - \boldsymbol{X}^*)\|_F^2 \\ &\quad= \|\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*) + \delta\mathcal{P}_\Omega(\boldsymbol{M}^* - \boldsymbol{M}^k)\|_F^2 \\ &\quad= \|\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*)\|_F^2 + \delta^2\|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2 \\ &\quad- 2\delta\langle\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*), \boldsymbol{M}^k - \boldsymbol{M}^*\rangle\end{aligned} \tag{36}$$

Combine (35) with (36),

$$\begin{aligned}\|\mathcal{P}_\Omega(\boldsymbol{X}^k - \boldsymbol{X}^*)\|_F^2 &\leqslant \|\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*))\|_F^2 \\ &+ \delta^2\|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2 - 2\delta\|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2\end{aligned} \tag{37}$$

Under the assumption that $0 < \delta < 2$, we have

$$2\delta - \delta^2 \geqslant \alpha \tag{38}$$

where some $\alpha > 0$, thus

$$\begin{aligned}\|\mathcal{P}_\Omega(\boldsymbol{X}^k - \boldsymbol{X}^*)\|_F^2 &\leqslant \|\mathcal{P}_\Omega(\boldsymbol{X}^{k-1} - \boldsymbol{X}^*))\|_F^2 - \\ &\alpha\|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2.\end{aligned} \tag{39}$$

Therefore, $\|\mathcal{P}_\Omega(\boldsymbol{X}^k - \boldsymbol{X}^*)\|_F^2$ is non-increasing and finally converges to a limit. Meanwhile, $\|\boldsymbol{M}^k - \boldsymbol{M}^*\|_F^2 \to 0$ as $k \to \infty$. Theorem 1 is established.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Environmental Protection Department (EPD), the Hong Kong SAR Government, for making the air pollution data freely accessible. We would also like to thank Mr. Han Yang, PhD student of Electrical and Electronic Engineering, and HKU-Cambridge Clean Energy and Environment Research Platform (CEERP), and HKU-Cambridge AI to Advance Well-being and Society Research Platform (AI-WiSe), the University of Hong Kong, for providing the valuable air quality dataset.

REFERENCES

[1] J. R. Wolch, J. Byrne, and J. P. Newell, "Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'," Landscape and urban planning, vol. 125, pp. 234–244, 2014.

[2] H. Chen, J. C. Kwong, R. Copes, K. Tu, P. J. Villeneuve, A. Van Donkelaar, P. Hystad, R. V. Martin, B. J. Murray, B. Jessiman et al., "Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study," The Lancet, vol. 389, no. 10070, pp. 718–726, 2017.

[3] A. J. Cohen, M. Brauer, R. Burnett, H. R. Anderson, J. Frostad, K. Estep, K. Balakrishnan, B. Brunekreef, L. Dandona, R. Dandona et al., "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases study 2015," The Lancet, vol. 389, no. 10082, pp. 1907–1918, 2017.

[4] S. Liu, Y. Zhou, S. Liu, X. Chen, W. Zou, D. Zhao, X. Li, J. Pu, L. Huang, J. Chen et al., "Association between exposure to ambient particulate matter and chronic obstructive pulmonary disease: results from a cross-sectional study in China," Thorax, vol. 72, no. 9, pp. 788–795, 2017.

[5] T. Li, R. Hu, Z. Chen, Q. Li, S. Huang, Z. Zhu, and L.-F. Zhou, "Fine particulate matter (PM 2.5): The culprit for chronic lung diseases in China," Chronic diseases and translational medicine, vol. 4, no. 3, pp. 176–186, 2018.

[6] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. ACM, 2013, pp. 1436–1444.

[7] K. R. Anderson, E. L. Avol, S. A. Edwards, D. A. Shamoo, R.-C. Peng, W. S. Linn, and J. D. Hackney, "Controlled Exposures of Volunteers to Respirable Carbon and Sulf uric Acid Aerosols," Journal of the Air & Waste Management Association, vol. 42, no. 6, pp. 770–776, Jun. 1992.

[8] V. O. K. Li, Y. Han, J. C. K. Lam, Y. Zhu, and J. Bacon-Shone, "Air pollution and environmental injustice: Are the socially deprived exposed to more PM 2.5 pollution in Hong Kong?" Environmental Science & Policy, vol. 80, pp. 53–61, 2018.

[9] C. K. Chan and X. Yao, "Air pollution in mega cities in China," Atmospheric environment, vol. 42, no. 1, pp. 1–42, Jan. 2008.

[10] R. M. Harrison and J. Yin, "Particulate matter in the atmosphere: which particle properties are important for its effects on health?" Science of the total environment, vol. 249, no. 1, pp. 85–101, Apr. 2000.

[11] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," Atmospheric Environment, vol. 38, no. 18, pp. 2895–2907, Jun. 2004.

[12] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," Computer Networks, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[13] J. D. Allan, J. L. Jimenez, P. I. Williams, M. R. Alfarra, K. N. Bower, J. T. Jayne, H. Coe, and D. R. Worsnop, "Quantitative sampling using an Aerodyne Aerosol Mass Spectrometer 1. Techniques of data interpretation and error analysis," Journal of Geophysical Research: Atmospheres, vol. 108, no. D3, pp. 4090–4091, 2003.

[14] F. M. Cleveland, "Cyber security issues for advanced metering infrasttructure (AMI)," in IEEE Power and Energy Society General Meeting, 2008, pp. 1–5.

[15] G. J. McRae and J. H. Seinfeld, "Development of a second-generation mathematical model for urban air pollution—ii. evaluation of model performance," Atmospheric Environment (1967), vol. 17, no. 3, pp. 501–522, 1983.

[16] V. O. K. Li, J. C. K. Lam, Y. Chen, and J. Gu, "Deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data," in IEEE Global Communications Conference, Dec. 2017, pp. 1–6.

[17] Y. Yu, J. J. Q. Yu, V. O. K. Li, and J. C. K. Lam, "Low-rank singular value thresholding for recovering missing air quality data," in IEEE International Conference on Big Data, Boston, USA, Dec. 2017, pp. 508–513.

[18] J. Y. Zhu, C. Sun, and V. O. K. Li, "An extended spatio-temporal granger causality model for air quality estimation with heterogeneous urban big data," IEEE Transactions on Big Data, vol. 3, no. 3, pp. 307–319, Sep. 2017.

[19] F. WANG, S.-y. CHENG, M.-j. LI, and Q. FAN, "Optimizing bp networks by means of genetic algorithms in air pollution prediction [j]," Journal of Beijing University of Technology, vol. 9, p. 016, 2009.

[20] R. A. Harley, A. G. Russell, G. J. McRae, G. R. Cass, and J. H. Seinfeld, "Photochemical modeling of the Southern California air quality study," Environmental Science & Technology, vol. 27, no. 2, pp. 378–388, Feb. 1993.

[21] R. Kohn and C. F. Ansley, "Estimation, prediction, and interpolation for ARIMA models with missing data," Journal of the American Statistical Association, vol. 81, no. 395, pp. 751–761, Sep. 1986.

[22] A. C. Kokaram, R. D. Morris, W. J. Fitzgerald, and P. J. Rayner, "Detection of missing data in image sequences," IEEE Transactions on Image Processing, vol. 4, no. 11, pp. 1496–1508, Nov. 1995.

[23] R. Lguensat, P. Tandeo, P. Ailliot, B. Chapron, and R. Fablet, "Using archived datasets for missing data interpolation in ocean remote sensing observation series," in OCEANS 2016-Shanghai. IEEE, Apr. 2016, pp. 1–5.

[24] M. N. Norazian, Y. A. Shukri, R. N. Azam, and A. M. M. Al Bakri, "Estimation of missing values in air pollution data using single imputation techniques," ScienceAsia, vol. 34, no. 3, pp. 341–345, Sep. 2008.

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

[26] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," Multiscale Modeling & Simulation, vol. 7, no. 3, pp. 1005–1028, Nov. 2008.

[27] J. Li, Q. Yuan, H. Shen, and L. Zhang, "Hyperspectral image recovery employing a multidimensional nonlocal total variation model," Signal Processing, vol. 111, pp. 230–248, Jun. 2015.

[28] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D: Nonlinear Phenomena, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.

[29] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 10, pp. 3660–3677, Oct. 2012.

[30] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[31] J. Li, Q. Yuan, H. Shen, and L. Zhang, "Noise removal from hyperspectral image with joint spectral–spatial distributed sparse representation," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 9, pp. 5425–5439, Sep. 2016.

[32] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral image restoration using low-rank matrix recovery," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 8, pp. 4729–4743, Aug. 2014.

[33] W. He, H. Zhang, L. Zhang, and H. Shen, "Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 1, pp. 178–188, Jan. 2016.

[34] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[35] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," IEEE Transactions on Information Theory, vol. 57, no. 3, pp. 1548–1566, 2011.

[36] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," SIAM Journal on Matrix Analysis and Applications, vol. 31, no. 3, pp. 1235–1256, Nov. 2009.

[37] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

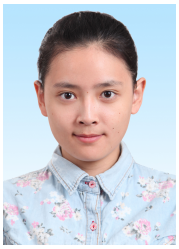
[38] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," Journal of machine learning research, vol. 11, no. Aug, pp. 2287–2322, Aug. 2010.

[39] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix factorization with missing data," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 2005, pp. 316–322.

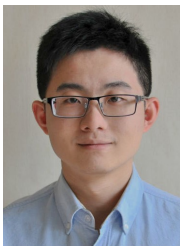
[40] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in Artificial Intelligence, vol. 2009, p. 4, 2009.

[41] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," IEEE Transactions on Power Systems, vol. 31, no. 2, pp. 1006–1013, Mar. 2016.

[42] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Transactions on Antennas and Propagation, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[43] S. Thirukumaran and A. Sumathi, "Missing value imputation techniques depth survey and an imputation algorithm to improve the efficiency of imputation," in 2012 Fourth International Conference on Advanced Computing (ICoAC). IEEE, 2012, pp. 1–5.
[44] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, 2011.
[45] L. F. Burgette and J. P. Reiter, "Multiple imputation for missing data via sequential regression trees," American journal of epidemiology, vol. 172, no. 9, pp. 1070–1076, 2010.
[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–22, 1977.
[47] P. Lodder, "To impute or not impute: That's the question," Advising on research methods: Selected topics. Huizen: Johannes van Kessel Publishing, 2013.
[48] P. Royston, I. R. White et al., "Multiple imputation by chained equations (mice): implementation in stata," J Stat Softw, vol. 45, no. 4, pp. 1–20, 2011.
[49] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?" International journal of methods in psychiatric research, vol. 20, no. 1, pp. 40–49, 2011.
[50] Air pollution Dataset. Accessed on: Jul. 10, 2017. [Online]. Available: https://cd.epic.epd.gov.hk/EPICDI/air/station/.
[51] Meterology Dataset. Accessed on: Jul. 10, 2017. [Online]. Available: http://www.hko.gov.hk/contentc.htm.
[52] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: a review," Atmospheric Environment, vol. 37, no. 2, pp. 155–182, Jan. 2003.
[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[54] M. R. Espejo, "The Oxford dictionary of statistical terms," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 167, no. 2, pp. 377–377, May. 2004.
[55] C. Coronel and S. Morris, Database systems: design, implementation, & management. Cengage Learning, 2016.
[56] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," Foundations of Computational mathematics, vol. 9, no. 6, p. 717, Dec. 2009.
[57] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in A Practical Approach to Microarray Data Analysis, Boston, MA, 2003, pp. 91–109.
[58] L. Chen, Y. Liu, and C. Zhu, "Iterative block tensor singular value thresholding for extraction of low rank component of image data," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE, Mar. 2017, pp. 1862–1866.
[59] D. P. Bertsekas and A. Scientific, Convex optimization algorithms. Athena Scientific Belmont, 2015.
[60] U. Langer and W. Queck, "On the convergence factor of Uzawa's algorithm," Journal of computational and applied mathematics, vol. 15, no. 2, pp. 191–202, Jun. 1986.
[61] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2053–2080, May, 2010.
[62] B. Recht, "A simpler approach to matrix completion," Journal of Machine Learning Research, vol. 12, no. Dec, pp. 3413–3430, Dec. 2011.
[63] G. H. Goluh and C. F. Van Loan, "Matrix computations," Johns Hopkins, 1996.
[64] T.-H. Oh, Y. Matsushita, Y.-W. Tai, and I. So Kweon, "Fast randomized singular value thresholding for nuclear norm minimization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4484–4493.
[65] Y. Li and W. Yu, "A fast implementation of singular value thresholding algorithm using recycling rank revealing randomized singular value decomposition," arXiv preprint arXiv:1704.05528, 2017.
[66] M. W. Kenyhercz and N. V. Passalacqua, "Missing data imputation methods and their performance with biodistance analyses," in Biological Distance Analysis. Elsevier, 2016, pp. 181–194.
[67] M. Slater, "Lagrange multipliers revisited," in Traces and Emergence of Nonlinear Programming 2014, Birkhäuser, Basel, 2014, pp. 293–306.
[68] S. Boyd and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.

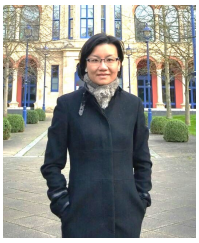


YANGWEN YU received the B.Eng. degree in Communication Engineering (Science Experimental Class) and the M.Eng. degree in Telecommunication & Information System from Beijing Jiaotong University, Beijing, China, in 2013 and 2016, respectively. She is currently pursuing a Ph.D. degree in the Department of Electrical & Electronic Engineering, the University of Hong Kong (HKU). Her research interests include information theory, coding and modulation, and temporally- and spatially-correlated missing data recovery.



JAMES J.Q. YU (S'11–M'15) received the B.Eng. and Ph.D. degree in the Department of Electrical and Electronic Engineering from the University of Hong Kong, Pokfulam, Hong Kong, in 2011 and 2015, respectively. He was a postdoctoral fellow at the University of Hong Kong from 2015 to 2018. He is currently an assistant professor at the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China, and an honorary assistant professor at the Department of Electrical and Electronic Engineering, the University of Hong Kong. He is also the chief research consultant of GWGrid Inc., Zhuhai, and Fano Labs, Hong Kong. His research interests include smart city technologies, deep learning and big data, intelligent transportation systems, and energy systems. He is an associate editor of IET Smart Cities.

VICTOR O.K. LI (S'80 – M'81 – F'92) received SB, SM, EE and ScD degrees in Electrical Engineering and Computer Science from MIT. Prof. Li is Chair of Information Engineering and Cheng Yu-Tung Professor in Sustainable Development at the Department of Electrical & Electronic Engineering (EEE) at the University of Hong Kong. He was Visiting Professor at the Department of Computer Science and Technology, University of Cambridge, UK, from April to August 2019. He is the Director of the HKU-Cambridge Clean Energy and Environment Research Platform, and of the HKU-Cambridge AI to Advance Well-being and Society Research Platform, which are interdisciplinary collaborations with Cambridge University. He was the Head of EEE, Assoc. Dean (Research) of Engineering and Managing Director of Versitech Ltd. He serves on the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange and co-founded Fano Labs Ltd., an artificial intelligence (AI) company with his Ph.D. student. Previously, he was Professor of Electrical Engineering at the University of Southern California (USC), Los Angeles, California, USA, and Director of the USC Communication Sciences Institute. His research interests include big data, AI, optimization techniques, and interdisciplinary clean energy and environment studies. In Jan 2018, he was awarded a USD 6.3M RGC Theme-based Research Project to develop deep learning techniques for personalized and smart air pollution monitoring and health management. Sought by government, industry, and academic organizations, he has lectured and consulted extensively internationally. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the HKSAR. He is a Fellow of the Hong Kong Academy of Engineering Sciences, the IEEE, the IAE, and the HKIE.



JACQUELINE C.K. LAM is Associate Professor of the Department of Electrical and Electronic Engineering, the University of Hong Kong, also Co-Director of the HKU AI to Advance Well-being and Society Institute, which embeds the HKU-Cambridge Clean Energy and Environment Research Platform, and the HKU-Cambridge AI to Advance Well-being and Society Research Platform. She has been the Hughes Hall Visiting Fellow before she takes up Visiting Senior Research Fellow and Associate Researcher in Energy Policy Research Group, Judge Business School, the University of Cambridge. She is currently both a visiting scholar at CEEPR, MIT and the Department of Computer Science and Technology, the University of Cambridge. Her research examines air pollution, and public health, using big data and deep learning techniques. Her work has been published in IEEE, Environment International, Applied Energy, Environmental Science and Policy, and Energy Policy. Jacqueline has received three times the research grants awarded by the Research Grants Council, HKSAR Government, from 2011-2017. The funded amount totaled USD 7.8M in PI or Co-PI capacity. Her recent research study, in collaboration with Yang Han and Victor OK Li on PM2.5 pollution and environmental inequality in Hong Kong, has been published in Environmental Science and Policy, and widely covered by more than 30 local and overseas newspapers and TVs. She recently co-organized a seminar "AI for Social Good" with Prof. Victor Li and Prof. Jon Crowcroft in Cambridge, which was well received.