Low-rank Singular Value Thresholding for Recovering Missing Air Quality Data

Yangwen Yu, James J.Q. Yu, Victor O.K. Li, and Jacqueline C.K. Lam
Department of Electrical and Electronics Engineering
The University of Hong Kong
Pokfulam, Hong Kong, China
Email: {ywyu,jqyu,vli,jcklam}@eee.hku.hk

***Abstract*—With the increasing awareness of the harmful impacts of urban air pollution, air quality monitoring stations have been deployed in many metropolitan areas. These stations provide air quality data to the public. However, due to sampling device failures and data processing errors, missing data in air quality measurements is common. Data integrity becomes a critical challenge when such data are employed for public services. In this paper, we investigate the mathematical property of air quality measurements, and attempt to recover the missing data. First, we empirically study the low rank property of these measurements. Second, we formulate the low rank matrix completion (LRMC) optimization problem to reconstruct the missing air quality data. The problem is transformed using duality theory, and singular value thresholding (SVT) is employed to develop sub-optimal solutions. Third, to evaluate the performance of our methodology, we conduct a series of case studies including different types of missing data patterns. The simulation results demonstrate that the proposed SVT methodology can effectively recover missing air quality data, and outperform the existing Interpolation. Finally, we investigate the parameter sensitivity of SVT. Our study can serve as a guideline for missing data recovery in the real world.**

***Keywords*-Missing data recovery, air quality measurements, low-rank matrix completion, singular value thresholding.**

I. Introduction

Air pollution presents a critical environmental challenge in modern cites [1]. Due to its adverse impacts on human health, citizens have a strong demand for timely reports of air quality measurements, especially $\mathrm{PM}_{2.5}$ and PM_{10}[1]. Air quality data, which in this paper refer to both air pollutants and meteorology data, are collected so as to provide public services such as real-time health alert and advice. However, the missing data problem affects the quality and utility of these data [1]. There are generally three factors contributing to data loss, namely, communication failures, facility faults, and cyber-security attacks.

Firstly, the local measured data from these monitoring stations need to be transmitted to a control center, and the communication infrastructure connecting these parties may suffer from random failures [2]. Secondly, the monitoring facilities installed in these stations may sometimes experience operational faults, which can lead to missing data for a period of time until fault recovery. Thirdly, as air quality monitoring based on the measured real-time data carries significant social impacts, the whole system is exposed to cyber-security attacks, where adversaries attempt to modify or erase the measurements [3].

For data loss caused by communication failures, missing data tend to be purely random, i.e., no temporal or spatial correlation. For the next two cases, missing data are more likely to be temporally related [1]. To the best of our knowledge, there is no dedicated solution to address both these two kinds of data loss problems. The conventional method to handle the missing data is merely by padding these entries with zeros or historical data interpolations, which can be either meaningless or inaccurate. This calls for the research community to develop more advanced information engineering techniques to recover the missing data and improve public services [4].

In this paper we employ a recent development in low-rank matrix completion (LRMC) to address the missing air quality data problem. LRMC is one of the variants of matrix completion. It can fill in the missing entries of a partial low rank matrix accurately. So this method has been applied to solve problems in various engineering research fields, such as social network [5] and power system [6]. As the air quality data share similar characteristics with data in these applications, it is possible that similar methods can be adopted to address the problem. We would like to point out that this paper is the first work that focuses on recovering missing air quality data. Our key contributions to the field of missing data recovery are as follows:

- We empirically analyze the matrix rank properties of air quality data, and suggest that low-rank matrix completion is a potential technique to solve the missing data problem.
- We formulate the missing air quality data recovery problem, and adopt Singular Value Thresholding (SVT) to develop sub-optimal solutions.
- We perform a series of comprehensive simulations to test the performance of SVT in addressing the missing data problem, and compare it with the existing Inter-

[1]PM stands for Particulate Matter. The subscript indicates the diameter of the particulate in microns. $\mathrm{PM}_{2.5}$ and PM_{10} are two of the most health-threatening air pollutants.

polation method.
- We investigate the parameter sensitivity of SVT on the data recovery accuracy, and summarize the best parameter values for real-world application.

The rest of this paper is organized as follows. Section II introduces the background of air quality data and the system model for missing air quality data recovery problem. Section III analyzes the low-rank property of air quality data through empirical tests. Section IV performs preliminary analyses on the proposed problem, and adopts the SVT technique to solve the transformed optimization problem. In Section V, the performance of SVT in the proposed missing data problem is assessed, and the parameter sensitivity of SVT is investigated. Finally, we concluded our work in Section VI, with discussions on our future.

II. MATHEMATICAL FRAMEWORK

A. Air quality measurement data

In this paper, 11 categories of air quality data have been retrieved for our study. These data are retrieved online from AQICN database, which collects air pollutants data from air quality monitoring stations and meteorology data from the observatories in Beijing, China [7]. As shown in Table I, these measurements include five categories of meteorological data and six categories of pollutants that are used to generate the Air Quality Index (AQI) in China. The air quality data are collected once per hour, and reported to the control center immediately. The data is from the first week of January 2017, i.e. Jan. 1-7, 2017.

Let $\boldsymbol{M}_{\mathrm{GT}} \in \mathbb{R}^{n_1 \times n_2}$ denote the ground truth data of $n_{21} = 11$ categories from n_{22} monitoring stations ($n_{21} \times n_{22} = n_2$) in n_1 time intervals. Ideally, these data can be applied in further air quality analyses and services, see [8] for example. However, in real life, these data often suffer from losses due to different reasons described in Section I. Let $\boldsymbol{M}_{\mathrm{OB}}$ be the data matrix available to the control center. $\boldsymbol{M}_{\mathrm{OB}}$ may have missing entries, which are replaced by zeros. We use Ω to denote the set of the indices (i, j) of observed data in $\boldsymbol{M}_{\mathrm{OB}}$. Then $\boldsymbol{M}_{\mathrm{OB}}$ and $\boldsymbol{M}_{\mathrm{GT}}$ satisfy the following relationship:

$$\boldsymbol{M}_{\mathrm{OB}}(i,j) = \begin{cases} \boldsymbol{M}_{\mathrm{GT}}(i,j) & (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $\boldsymbol{M}_{\mathrm{OB}}(i,j)$ and $\boldsymbol{M}_{\mathrm{GT}}(i,j)$ are the data located at entry (i,j) of matrices $\boldsymbol{M}_{\mathrm{OB}}$ and $\boldsymbol{M}_{\mathrm{GT}}$, respectively. For simplicity, equation (1) can be expressed as $\boldsymbol{M}_{\mathrm{OB}} = P_{\Omega}(\boldsymbol{M}_{\mathrm{GT}})$, where $P_{\Omega}(\cdot)$ is a projection $\mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$.

B. Missing data recovery model

We attempt to recover the missing data entries in matrix $\boldsymbol{M}_{\mathrm{OB}}$ with the observed data entries. This problem is defined as the Air Quality Data Recovery (AQDR) problem, which can be formulated as follows:

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

where $\boldsymbol{M}$ is the recovered air quality data matrix. This matrix should keep all observed data while minimizing the difference from the ground truth $\boldsymbol{M}_{\mathrm{GT}}$ on missing data entries. Ideally, the optimal recovered matrix is identical to the ground truth, and acts as a reliable data source for subsequent analyses.

In practice, however, the ground truth $\boldsymbol{M}_{\mathrm{GT}}$ is only partially known in the form of $\boldsymbol{M}_{\mathrm{OB}}$. Considering that $\boldsymbol{M}_{\mathrm{GT}}$ in the objective function (2a) is unknown when conducting the optimization, the single constraint (2b) cannot ensure a unique solution to the problem can be derived. In addition, as the constraint does not impose any limitations on the non-observed entries, the solutions to (1) cannot be considered a significant approximation.

In the past decade, research has been conducted on re-formulating the matrix completion problem as a convex optimization problem by introducing an extra regularization term to obtain a stable and accurate estimated solution [9]. The matrix completion problem shown in (2) can be converted to the following well-posed form:

$$\text{minimize } R(\boldsymbol{M}) \tag{3a}$$
$$\text{subject to } P_{\Omega}(\boldsymbol{M}) = P_{\Omega}(\boldsymbol{M}_{\mathrm{GT}}) \tag{3b}$$

In problem (3a), $R(\boldsymbol{M})$ is the regularization term of $\boldsymbol{M}$. This regularization can be the norms of $\boldsymbol{M}$ that are related to some specific properties of the matrix such as low-rank and high sparsity. By integrating the characteristics of the data matrix $\boldsymbol{M}$, (3a) can obtain a unique optimal solution $\boldsymbol{M}^*$ that accurately resembles the ground truth matrix $\boldsymbol{M}_{\mathrm{GT}}$.

As illustrated in Section III, the investigated air quality data matrix observes the low-rank property. Hence, we employ the low-rank model to develop the regularization term for the AQDR problem.

Table I
DATA COLLECTED

Domain	Category	Data Source
Air pollutant	$PM_{2.5}$	AQICN [7]
	PM_{10}	
	O_3	
	NO_2	
	SO_2	
	CO	
Meteorology	Temperature (°C)	
	Dew Point (°C)	
	Pressure (Pa)	
	Humidity (%)	
	Wind (m/s)	

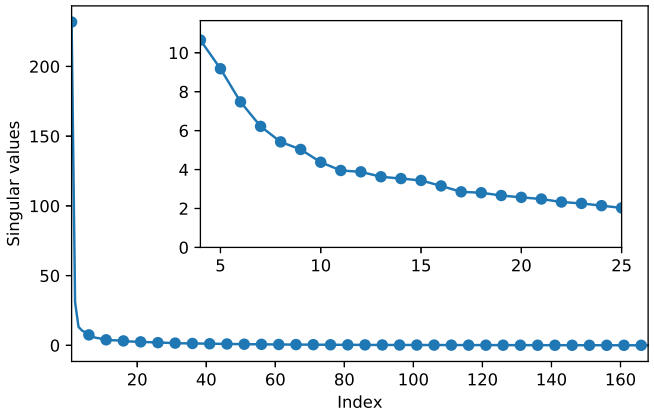

Figure 1. Distribution of singular values of an arbitrary air quality data matrix $\boldsymbol{M}_{GT}$.

Table II
THE LARGEST 45 SINGULAR VALUES OF $\boldsymbol{M}_{GT}$

	σ_i				
rank	1∼9	10∼18	19∼27	28∼36	37∼45
1	231.8317	4.3756	2.6669	1.7681	1.2869
2	31.0327	3.9525	2.5719	1.7517	1.2707
3	13.1812	3.8824	2.4864	1.6889	1.1985
4	10.6508	3.6296	2.3318	1.5907	1.1596
5	9.1780	3.5365	2.2601	1.5026	1.1322
6	7.4756	3.4416	2.1456	1.4738	1.1191
7	6.2178	3.1591	2.0293	1.4372	1.1001
8	5.4208	2.8555	1.9134	1.3816	1.0689
9	5.0328	2.8094	1.8670	1.3305	0.9839

III. Rank Analysis of Air Quality Data

The "rank" of a matrix refers to the maximum number of linear-independent columns [10], which can be interpreted as the degree of redundancy in air quality data. Consequently, the rank of the air quality data matrix M_{GT} is notably smaller than its dimensions, i.e. n_1 and n_2, as illustrated in the later part of this section. Such matrices with small rank values are considered as "low-rank" [9].

Generally speaking, the air quality data is dynamic over the long run for an arbitrary region. At the same time, the short-term variation of each category within several hours can be very limited. Moreover, according to the micro-scale dispersion model [8], the concentration of air pollution demonstrates significant spatio-temporal (S-T) dependency and can be drastically influenced by urban meteorology. Therefore, it is possible that the air quality data matrix has high data redundancy and has strong correlation among its entries. The air quality matrix thus displays the low-rank property.

In order to determine the rank of the target data matrix, we analyze the distribution of singular values for a single data matrix. The investigated data matrix of the ground truth $\boldsymbol{M}_{\mathrm{GT}} \in \mathbb{R}^{168\times 330}$ contains 11 categories of data collected from 30 locations distributed across the city of Beijing in the first week of January 2017. Similar to other data processing methods [11], [12], we normalize the measurement matrix in order to improve the effectiveness of data processing.

Utilizing the singular value decomposition (SVD) method [13], the singular values of $\boldsymbol{M}_{GT} \in \mathbb{R}^{168\times 330}$ can be computed as follows:

$$\boldsymbol{M}_{GT} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{4}$$

where $\boldsymbol{U} \in \mathbb{R}^{168\times 168}$ and $\boldsymbol{V} \in \mathbb{R}^{330\times 330}$ are unitary matrices, $\boldsymbol{\Sigma}$ is a 168×330 rectangular diagonal matrix. The non-negative diagonal entries σ_i in Σ are the singular values of $\boldsymbol{M}_{GT}$, which are essential in analyzing the low-rank property of the target matrix.

As depicted in Fig. 1, the singular values of $\boldsymbol{M}_{GT}$ diminish quite quickly. This suggests that the matrix can be approximated by a low-rank matrix in high accuracy [14]. This low-rank property encourages the transformation of the AQDR problem into a LRMC problem. To solve the problem, an estimated matrix with a low rank is used to approximate the ground truth matrix. In this paper, we employ all singular values that are greater than one as suggested by [6]. As a result, the rank of the approximated matrix is 45. The selected singular values are listed in Table II.

IV. Singular Value Thresholding for Air Quality Data Recovery

LRMC was firstly proposed by Candés and Recht [9] to exactly recover data matrices with low ranks using partially observed entries. By introducing the regularization term of LRMC, problem (3a) can be expressed as the following convex minimization problem [9]:

$$\text{minimize } \|\boldsymbol{M}\|_* \tag{5a}$$
$$\text{subject to } P_\Omega(\boldsymbol{M}) = P_\Omega(\boldsymbol{M}_{\mathrm{GT}}) \tag{5b}$$

where the nuclear norm $\|\boldsymbol{M}\|_*$ is the sum of the singular values of $\boldsymbol{M}$. This problem is actually a tight convex relaxation of the rank minimization problem, which is NP-hard [15]:

$$\text{minimize } \mathrm{rank}(\boldsymbol{M}) \tag{6a}$$
$$\text{subject to } P_\Omega(\boldsymbol{M}) = P_\Omega(\boldsymbol{M}_{\mathrm{GT}}) \tag{6b}$$

since the ball $\{\boldsymbol{M}|\|\boldsymbol{M}\|_* \leq 1\}$ is the convex hull of the set of rank-1 matrices with spectral norms bounded by one [9]. Therefore, the optimal solution of (5) corresponds to the best approximation of the ground truth matrix $\boldsymbol{M}_{\mathrm{GT}}$. While (5) is convex, common optimization solvers do not support optimizing over matrix ranks. To solve this problem, SVT [16] is proposed as a sub-optimal solution to the original LRMC problem (5). SVT substitutes the objective function (5a) by the following modified function:

$$\text{minimize } \tau\|\boldsymbol{M}\|_* + \frac{1}{2}\|\boldsymbol{M}\|_F^2, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. Obviously, when $\tau \to \infty$, the solution of objective function (7) converges to that of (5a). According to [16], a sub-optimal solution of (5a) can be obtained from solving (7) with a relatively large τ, suggested as $\sqrt{n_1 n_2}$ in [16]. Based on the sub-gradient method, SVT can solve this modified optimization problem (7) in an iterative manner [16].

As mentioned in Section III, SVD for a matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank r is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. The core operator of the SVT algorithm, named singular value shrinkage, utilizes the values of $\boldsymbol{U}$, $\boldsymbol{V}$, and a truncated $\boldsymbol{\Sigma}$ to create a new matrix which can approximate $\boldsymbol{X}$:

$$D_\tau(\boldsymbol{X}) \equiv \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}_\tau \boldsymbol{V}^T, \tag{8}$$

where $\tilde{\boldsymbol{\Sigma}}_\tau = \text{diag}\{\sigma_i - \tau\}_+$, and operator $\{t\}_+ = max(0, t)$. With this operator, SVT approaches the optimal solution to (7) in an iterative manner. SVT first initializes an intermediate matrix $\boldsymbol{X}^0 = \boldsymbol{0}$. Then in subsequent iteration $k = 1, 2, \cdots$, this intermediate matrix as well as the recovered matrix $\boldsymbol{M}^k$ are updated using the following rules:

$$\boldsymbol{M}^k = D_\tau(\boldsymbol{X}^{k-1}), \tag{9a}$$
$$\boldsymbol{X}^k = \boldsymbol{X}^{k-1} + \delta P_\Omega(\boldsymbol{M}_{\text{OB}} - \boldsymbol{M}^k), \tag{9b}$$

In SVT, this calculation repeats until a termination criterion is met:

$$\frac{\|P_\Omega(\boldsymbol{M}^k - \boldsymbol{M}_{OB})\|_F}{\|P_\Omega(\boldsymbol{M}_{OB})\|_F} \leq \epsilon \tag{10}$$

where ϵ is the convergence threshold, which is typically set to a small positive value, e.g., 10^{-3}.

Moreover, the proper selection of step size δ is critical to achieve optimal convergence speed and data recovery accuracy. In [16], the authors suggested the following equation for δ:

$$\delta = 1.2\frac{n_1 n_2}{m} \tag{11}$$

where m is the number of observed entries in $\boldsymbol{M}_{\text{OB}}$. In addition, it is suggested that δ should always be selected in $[0, 2]$ for better convergence speed. The sensitivity of δ to the final data recovery accuracy will be investigated in Section V-C.

V. Case Studies

We conduct a series of simulations to demonstrate the efficacy of using SVT in recovering missing air quality data. We first test the data recovery accuracy of SVT in handling missing air quality data which are not S-T correlated. Then we investigate if SVT can yield similar performance when the missing data are temporally related, which is commonly found in real-world air quality data. Last but not the least, we assess the parameter sensitivity of step size δ of SVT on recovery accuracy, and suggest its potential best values.

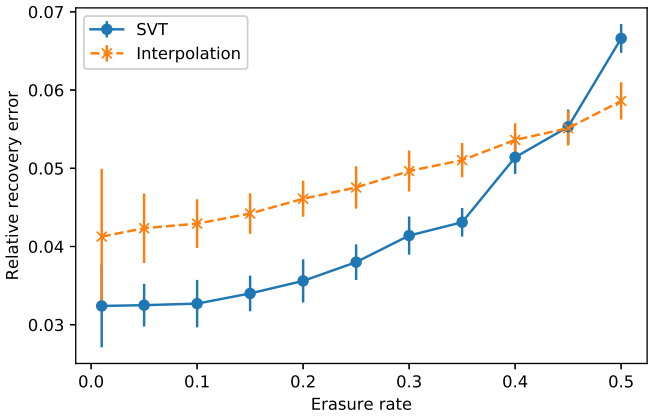

Figure 2. Relative recovery errors of SVT and Interpolation for random missing data entries.

The results of our case studies can be used as a guideline for practical implementations.

In these tests, we use the air quality data as described in Section III as the ground truth and erase some of the data. The recovery performance is measured by the relative recovery error

$$\text{Relative recovery error} = \frac{\|\boldsymbol{E} - \boldsymbol{E}_{\text{rec}}\|_F}{\|\boldsymbol{E}\|_F}, \tag{12}$$

where $\boldsymbol{E}$ represents the missing data entries in $\boldsymbol{M}_{\text{GT}}$, and $\boldsymbol{E}_{\text{rec}}$ represents the recovered measurements of these entries. Moreover, the relative recovery error is compared with Interpolation, in which, the missing entries in a matrix are recovered by the linear interpolates of their nearest non-missing neighbors.

A. Recovery of random missing data

We first evaluate the performance of SVT in recovering random missing data in air quality measurements. We randomly remove entries in $\boldsymbol{M}_{\text{GT}}$ to construct $\boldsymbol{M}_{\text{OB}}$. Each entry is removed with a probability equal to p, which is called the "erasure rate". Values for step size δ is set according to (11). All results are averaged based on 50 runs, in which the missing entries are randomly generated for each run. The simulation results and comparisons with Interpolation are depicted in Fig. 2, where both the averaged results and standard deviations are shown.

As illustrated in the figure, the recovery error increases with erasure rate p for both SVT and Interpolation. When $p < 0.4$, SVT significantly outperforms Interpolation. Meanwhile, when the erasure rate is extremely high ($p \geq 0.4$), Interpolation performs slightly better. However, it is highly unlikely that in reality the data may suffer from such a high loss rate. Therefore, it can be concluded that SVT generally achieves better missing data recovery performance as compared to Interpolation.

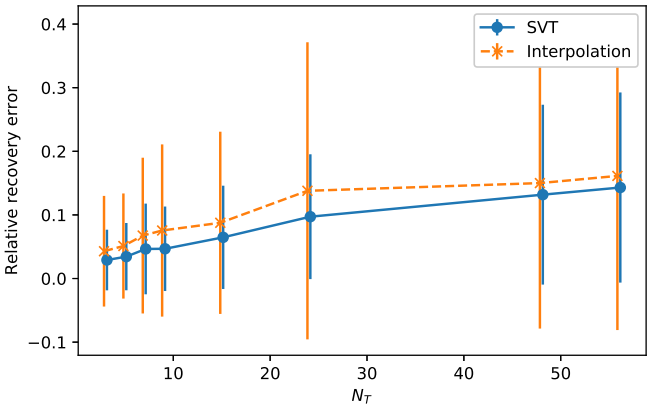

Figure 3. Relative recovery errors of SVT and Interpolation for time-correlated random missing air pollutant and meteorology data entries.

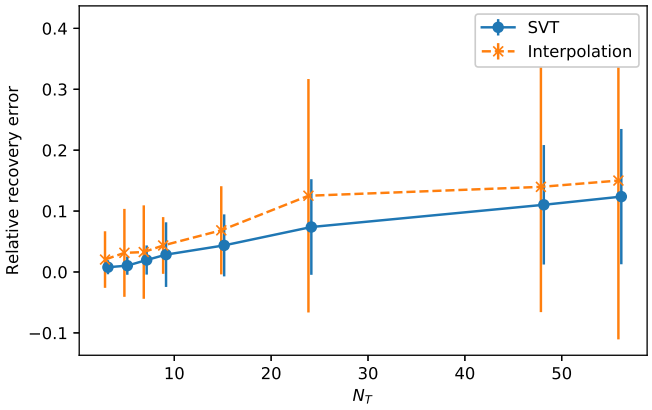

Figure 4. Relative recovery errors of SVT and Interpolation for time-correlated random missing air pollutant data entries.

Meanwhile, the performance degradation for SVT on $p \geq 0.4$ can be contributed by the selection of δ value. When $p = 0.4$, $\delta \approx 2.33$ according to (11), which exceeds the suggested upper limit for the step size suggested by [9]. Therefore, the convergence of the algorithm is potentially constrained, and the algorithm develops subpar recovery performance.

B. Recovery of temporal-correlated missing data

In the real world, it is common that several consecutive data are lost due to equipment failures and other issues. In such cases, the erasure of data are temporally correlated. Therefore, it is interesting to investigate the performance of SVT for recovering this kind of incomplete air quality data. In the empirical test, we develop the observed data matrix by removing consecutive samples. We use N_t to denote the number of temporally consecutive missing entries, and remove N_t consecutive entries in $\boldsymbol{M}_{\mathrm{GT}}$ randomly.

In order to focus on the performance of SVT in handling such cases, all other entries are considered available. In addition, for complete investigation, we consider cases where the erasure is either imposed randomly on all 11 categories of air pollutant and meteorology data, or on the six air pollutant data as shown in Table I. While the former generalizes the overall performance of SVT, the latter is more commonly seen in the real-world context. All simulations have been run 20 times for statistical significance. The averaged recovery errors and their standard deviations are presented in Figs. 3 and 4.

Considering both figures, it can be observed that SVT can always outperform Interpolation. This observation again demonstrates the superiority of SVT in recovering air quality data. In addition, even when a significant number of consecutive data are removed, e.g., $N_t > 40$, SVT can still partially recover the data, which will still be considered valuable as the recovered data serve as the only available estimation of the missing data.

C. Sensitivity of SVT Step Size δ

As found in the previous tests, the values for the step size of SVT δ is set according to (11), which is recommended by [16]. Meanwhile, for different erasure rates p, it is possible that changing the δ value can lead to a better data recovery accuracy than (11). Therefore, we perform a parameter sensitivity test for δ with respect to different p values. Specifically, (11) can be interpreted in the form $\delta = C(1-p)^{-1}$, where C is 1.2 in the equation for all previous tests. We test the performance of SVT with $C \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8\}$ and $p \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$. For each combination of C and p, 20 random runs are conducted. The averaged relative recovery errors and standard deviations (in brackets) are presented in Table III, where the best performing δ values are bolded.

From this table it can be concluded that SVT favors $C \geq 1.2$ when the erasure rate is small ($p \leq 0.1$), and the performance becomes sensitive to the selection of step size with the increase of p. In addition, when addressing problems with large p values, SVT favors medium C values more, and the best performing C decreases with larger p. As analyzed in Section V-A, this trend may be contributed by the drastically increasing δ values calculated based on p. To conclude, $C \in \{1.2, 1.4, 1.6\}$ can generally develop satisfactory missing data recovery error, and can be employed in real-world situations.

VI. Conclusion and Future Work

In this paper, we formulate a missing data recovery problem for air quality data, namely, the AQDR problem. Since the original problem is ill-posed, and an optimal solution cannot be easily obtained, the problem is transformed into a well-posed form. In addition, we investigate the low-rank property of the air quality data. The formulated problem is transformed into an LRMC equivalent problem for easy

Table III
RELATIONSHIP BETWEEN MISSING RATE p AND STEP SIZE δ

	δ							
p	0.01	0.05	0.1	0.15	0.2	0.3	0.4	0.5
$0.4(1-p)^{-1}$	0.039 (0.010)	0.039 (0.003)	0.040 (0.003)	0.042 (0.001)	0.043 (0.002)	0.049 (0.002)	0.057 (0.002)	0.068 (0.001)
$0.6(1-p)^{-1}$	0.035 (0.008)	0.035 (0.004)	0.037 (0.002)	0.038 (0.002)	0.039 (0.001)	0.045 (0.001)	0.055 (0.001)	0.066 (0.001)
$0.8(1-p)^{-1}$	0.032 (0.010)	0.033 (0.002)	0.034 (0.003)	0.036 (0.002)	0.036 (0.001)	0.043 (0.002)	0.052 (0.002)	**0.064 (0.001)**
$1.0(1-p)^{-1}$	0.033 (0.006)	0.033 (0.002)	0.032 (0.002)	0.034 (0.003)	0.036 (0.003)	0.042 (0.001)	0.052 (0.002)	**0.064 (0.001)**
$1.2(1-p)^{-1}$	0.032 (0.004)	0.032 (0.003)	0.032 (0.003)	0.034 (0.002)	0.035 (0.002)	**0.041 (0.001)**	**0.051 (0.001)**	0.066 (0.001)
$1.4(1-p)^{-1}$	0.032 (0.007)	0.033 (0.002)	0.033 (0.002)	**0.033 (0.002)**	0.034 (0.002)	0.041 (0.002)	0.054 (0.002)	0.074 (0.005)
$1.6(1-p)^{-1}$	0.032 (0.006)	**0.032 (0.002)**	0.032 (0.003)	0.033 (0.002)	**0.034 (0.001)**	0.050 (0.002)	0.075 (0.004)	0.094 (0.007)
$1.8(1-p)^{-1}$	**0.032 (0.006)**	0.032 (0.003)	**0.032 (0.002)**	0.050 (0.002)	0.069 (0.006)	0.102 (0.010)	0.136 (0.012)	0.175 (0.001)

solution. The relaxed problem is subsequently solved using SVT, a recently invented engineering technique to tackle the LRMC optimization problem.

We conduct comprehensive simulations to test the performance of SVT in addressing the formulated AQDR problem. The simulation results indicate that SVT outperforms the commonly used Interpolation in most test cases, which demonstrate that SVT is a practical method to address the missing air quality data problem. In addition, we investigate the impact of changing the step size value of SVT on the data recovery accuracy and suggest the best values for test cases of different characteristics. Results generated from our case studies can guide the adoption of SVT in real world missing data recovery. This method can also be extended to handle missing data problems in other research areas.

ACKNOWLEDGMENT

This research project is partially supported by the University Development Fund allocated for the HKU-Cambridge Clean Energy and Environment Research Platform (CEERP). The authors would like to gratefully acknowledge Mr. Han Yang, PhD student, Department of Electrical and Electronic Engineering, the University of Hong Kong, and CEERP, for providing the valuable air quality datasets to facilitate the empirical part of this study.

REFERENCES

[1] 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, 2004.

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

[3] F. M. Cleveland, "Cyber security issues for advanced metering infrasttructure (AMI)," in *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*. IEEE, 2008, pp. 1–5.

[4] 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.

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

[6] 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, 2016.

[7] Real-time Air Quality Index. Available: http://aqicn.org/city/beijing/.

[8] 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, 2003.

[9] E. J. Candés and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1007/s10208-009-9045-5

[10] G. Strang, *Introduction to linear algebra*, 5th ed. Wellesley-Cambridge Press Wellesley, MA, 2016.

[11] 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, 2004.

[12] C. Coronel and S. Morris, *Database systems: design, implementation, & management*. Cengage Learning, 2016.

[13] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*. Springer, 2003, pp. 91–109.

[14] L. Chen, Y. Liu, and C. Zhu, "Iterative block tensor singular value thresholding for extraction of low rank component of image data," *arXiv preprint arXiv:1701.04043*, 2017.

[15] G. J. Woeginger, "Exact algorithms for NP-hard problems: A survey," *Lecture Notes in Computer Science*, vol. 2570, no. 2003, pp. 185–207, 2003.

[16] 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, 2010.