Robust temporal low-rank representation for traffic data recovery via fused lasso

Ruyong Mao1 | Zhengyu Chen2 | Guobing Hu2

1 College of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
2 School of Electronic and Information Engineering, Jinling Institute of Technology, Nanjing, China

Abstract
Achieving complete and accurate traffic data as input is crucial for most intelligent transportation systems. However, due to hardware or software malfunction, traffic data is inevitably faced with missing and noise problems. Most of the existing representation-based traffic data recovery methods adopt sparse representation theory, which well models the local association properties of traffic data, but ignores their global correlation. To overcome this shortcoming, a robust low-rank representation method that incorporates temporal prior information to impute the missing traffic data is proposed. Specifically, the low-rank representation theory is first employed to model the global spatial correlation of traffic data, and then the fused lasso regularisation is utilized to fit the temporal correlation of traffic data. In addition, to make the proposed model more robust, F-norm regularisation is used to smooth the Gaussian noise of traffic data. Furthermore, an efficient optimisation algorithm based on ADMM is developed to solve the proposed model. Finally, the extensive experiments performed on real dataset validate the effectiveness of the proposed method.

1 INTRODUCTION

With the increase of car ownership year by year, urban roads have become more congested and the traffic efficiency is lower. In this context, Intelligent Transportation Systems (ITS) play a huge role and have received more and more research attention. [1, 2] Complete and accurate historical traffic data is the basis of ITS, but in actual conditions, the data collected by the traffic sensors suffers missing or corrupted value problem is almost inevitable. For example, in Melbourne, Australia, about 8% of the traffic flow sensor has a data loss rate as high as more than 56%, while in Beijing, China, about 10% of traffic data is incorrect. [3] There are many reasons for the above matter, mainly including storage device failure, transmission failure, data abnormality, and so forth. [4] However, most ITS must take complete and precise traffic data as input, such as short-term traffic flow predictions, [5, 6] so how to recover missing data with high accuracy has become a great challenge.

In order to address the problem of missing data mentioned above, scholars have proposed solutions from different perspectives based on the study of the characteristics of traffic data, among which the most representative is the Mean Imputation (MI), [7] K-nearest Neighbour Regression (KNNR), [8] Probabilistic Principal Component Analysis (PPCA), [9] Singular Value Threshold algorithm (SVT), [10] and so on. These methods have achieved good results, and some algorithms even have landmark significance. However, some of these proposed methods could not make full use of the spatial-temporal correlation of traffic flow data, moreover, some did not consider the impact of noise on missing value estimation. Considering the low-rank property of traffic flow data, the current advanced Low-rank Representation theory (LRR) can be used for modelling. In addition, low-rank representation also has the advantage of noise robust. Further, through an in-depth analysis of real traffic flow data, it is found that the traffic flow data has temporal stability property, that is, the traffic flow value changes between two neighbouring moments very small or even unchanged, so fused lasso [11] can be adopted to increase the precision of recovery.

Inspired by the above discussion, in this article we propose a novel noise robust traffic flow estimation model based on LRR. The contributions of this article are summarized as follows:
• By analysing a large amount of real traffic flow data, the hypothesis that the traffic flow matrix has a low rank is verified, and further analysis finds that the traffic flow data has temporal stability property.
• Based on the theory of LRR, and taking full advantage of the low-rank and temporal property of the traffic flow matrix, a noise robust completion model for missing traffic data with a temporal regularisation constraint term is proposed, named RTLRR (Robust Temporal Low-rank Representation).
• An optimized algorithm for solving the proposed model is designed based on the ADMM algorithm. This algorithm has the advantage of fast convergence.
• Based on the pre-processed raw dataset, the superiority of the proposed model compared to other methods is verified through comprehensive simulation experiments.

The organisational structure of the article is as follows: In Section 2, related works will be introduced. Section 3 will show our verification of prior hypothesis information on traffic flow data. In Section 4, we will propose a novel traffic flow missing data recovery model and give an optimisation algorithm. Then, in Section 5, we will compare the estimation results of the proposed model with other methods on outlier-free and raw datasets by simulation. Finally, the research work of this article will be summarized in Section 6.

2 | RELATED WORK

So far, the methods for recovering missing data can be roughly divided into the following three categories.

2.1 | Methods based on statistical learning

Statistical learning methods include simple interpolation methods, regression methods, and probabilistic model methods. The typical representative of naive interpolation method is MI. MI is to replace the missing values with the average of the observed samples. Tak et al. used such a method. [7] The advantage of MI is particularly simple, but because all missing values of the same sample use the same mean. Instead, the original distribution of the sample is seriously distorted, and the performance is often poor.

The typical representative of naive interpolation method is MI. MI is to replace the missing values with the average of the observed samples. Tak et al. used such a method. [7] The advantage of MI is particularly simple, but because all missing values of the same sample use the same mean. Instead, the original distribution of the sample is seriously distorted, and the performance is often poor.

The goal of regression is to establish a mapping relationship model between missing values and estimated values. Using such mapping relationships can easily estimate missing values. Regression can be divided into linear regression and non-linear regression. The representative of linear regression model is Least Square Regression (LSR). Non-linear regression mainly includes KNNR, Support Vector Regression (SVR) [12] and so on. The KNN algorithm finds the k samples closest to it based on the complete variables of missing samples in the data, which has been proved to have achieved good results in traffic flow data recovery. For example, Batista et al. proposed a data-driven improved KNN based regression algorithm called KNNR to estimate missing traffic values which achieved good results. [8] However, KNN recovers missing data in isolation, so it does not capture the global characteristics of the data. In addition, KNN has the disadvantage of being sensitive to noise. Shang et al. [12] proposed a traffic data estimation model based on Fuzzy Clustering Method (FCM) and SVR, and designed a solution method based on Particle Swarm Optimisation (PSO). This model not only uses the temporal characteristics of traffic flow data, but also takes into account the correlation between different sensors, so it has achieved good results. However, the model does not consider the noise factor in the flow data.

The probabilistic model method is based on the assumption that the data follow a certain probability distribution, such as uniform distribution, Gaussian distribution, chi-square distribution, and so forth, and then the missing values are estimated according to the probability model. PPCA is the representative of the probabilistic model method, Batista et al. [9] used this method for modelling. The model assumes that traffic data $x$ is generated by the hidden variable $z$, $x$ and $z$ are mapped as the relationship $x = Wz + \mu$, and the edge distribution $p(z)$ and conditional probability $p(x|z)$ of $z$ both obey Gaussian distribution. In such a linear Gaussian model, $p(x)$ and $p(z|x)$ also obey Gaussian distribution. The model optimisation goals are $W$ and $\mu$. Based on the prior assumptions, the model parameters can be optimized by using Maximum Likelihood Error (MLE) or Expectation Maximisation (EM). Recently, Bayesian Optimisation (BO) has also been widely used to adjust parameters. [13, 14] The disadvantage of PPCA is that it relies too much on the prior assumptions. However, in actual situations, it is difficult to determine which distribution the true data obeys. For example, traffic flow data often does not obey the normal distribution for the double peak traffic period.

2.2 | Methods based on matrix completion

Matrix completion theory is derived from Compressed Sensing Theory (CS). [15] Candès et al. [16] proved that the incomplete matrix can be recovered with higher accuracy only if the matrix to be restored is low-rank. For traffic flow matrix, this assumption is reasonable, because each item in the same road network are often related, which is called spatial correlation. Spatial correlation provides a basis for our LRR modelling. The traffic flow values at different times on the same road are also related, which is called temporal correlation. In general, the temporal correlation is represented by temporal stability, that is, the values at adjacent moments do not change much, and the temporal-spatial correlation is reflected by the low-rank property of the matrix. This will be demonstrated experimentally in the next Section 3. Commonly used matrix completion algorithms include SVT, OptSpace, [17] Subspace Evolution and
Transfer (SET), [18] and so on. In, [19] Luo et al. proposed an improved low rank matrix decomposition (ILRMD), which fully utilizes the spatiotemporal correlation characteristics among traffic data, but it could not care the impact of stochastic fluctuation of traffic data on estimation results. In recent years, research interest has gradually shifted from matrix completion in two-dimensional space to tensor completion in three-dimensional space. [20] proposed a flow estimation model combining matrix completion and tensor completion, and designed a flow estimation model. Based on the optimisation model of CP decomposition, the experimental results show that the method has a fairly high estimation accuracy, and it still has a good performance even on holidays with complex traffic conditions. However, this method cannot concern the local properties of samples and is sensitive to noise.

2.3 Methods based on sparse representation

In recent years, Sparse Representation theory (SR) [21] has been applied to the restoration of missing traffic flow data. Chen et al. [22] proposed a missing traffic value interpolation model based on $l_p$ norm regularisation called SSRp. In the theory SR, each sample can be represented by a linear combination of other samples. The goal is to find such a sparse coefficient matrix, which is different from the 1 norm constraint used in traditional sparse representation. [22] adopted constrained by the $l_p$ norm, a more sparsity was achieved. The simulation results show that the model has a high accuracy estimation performance. However, sparse representations have the inherent disadvantages of not being able to make good use of global properties and poor anti-noise performance. What is more, SSRp does not take into account the temporal stability characteristics of traffic flow.

Table 1 shows the summary of the related works mentioned in this section clearly.

| Methods based on               | Examples                                      |
|-------------------------------|------------------------------------------------|
| Statistical learning          | Simple interpolation methods (MI[7])          |
|                               | Regression methods (KNNR[8], SVR[12])         |
|                               | Probabilistic model methods (PPCA[9], BO-PPCA[13]) |
| Matrix completion             | SVT[10], OptSpace[17], SET[18], ILRMD[19]     |
| Sparse representation         | SSRp[22]                                       |

3 RESEARCH ON REAL TRAFFIC FLOW DATA

The real traffic flow data set used in this study is derived from 20 sensor nodes in the road network formed by the intersection of the I205 Freeway and SR14 Highway in Portland, Oregon, USA (http://portal.its.pdx.edu/). The map of the road network and the locations of 20 detection points are shown in Figure 1. Among them, there are 10 detection points on the SR14 road and 10 detection points on the I205 road. The detection point on the south side of the SR14 highway detects the traffic flow from west to east, and the opposite on the north side; the I205 detection point on the west side detects the traffic flow from north to south, and the opposite on the east. The collection time was from 1 July 2018 to 30 September 2018. Excluding holidays and weekends, 63 days of traffic flow data were obtained. The sampling resolution of the traffic flow is 5 min. It can be seen that there are 288 sampling points per day, and the total data amount is $288 \times 63 \times 20 = 362,880$.

To visually show the characteristics of traffic flow data, we select four typical figures in Figure 2 which is a visualisation of traffic data collected by 4 different sensors in the same day. The horizontal axis is time, and the vertical axis is vehicle flow. It is obvious that traffic data collected by the 20 sensors show some
comm on characteristics. First, the traffic data collected by most sensors have two peak periods in the morning and evening, and the two peak periods appear basically the same, that is, traffic flow is periodic. Secondly, the change of the traffic values at most times is relatively small, but at some moments a sudden change (surge and decrease) will occur.

3.1 | Low-rank property

As mentioned earlier, the spatio-temporal correlation is manifested by the low rank of the matrix. Low-rank property of the matrix is the basis for applying LRR. In this section, we will experimentally verify that the traffic flow data does have low rank property. We will use Singular Value Decomposition (SVD) to evaluate whether the traffic matrix is low-rank. Any traffic matrix $X \in R^{T \times N}$, where $T$ is the number of time slots and $N$ is the number of traffic flow detectors set at different location, $X$ can be decomposed into the following form

$$X = U \Sigma V^T$$ (1)

where $U \in R^{T \times T}$ and $V \in R^{N \times N}$ are two unitary matrices, and $\sigma \in R^{T \times N}$ is a diagonal matrix whose diagonal elements are singular values. The singular value of $X$ (in descending order) is $\text{diag}(\Sigma) = \{\sigma_1, \sigma_2, ..., \sigma_r, 0, ..., 0\}$, where $r$ is the rank of the matrix, obviously the rank of the matrix can use the number of non-zero singular value to represent. The definition of the rank of the above matrix is actually the exact rank of the matrix. It is worth noting that in the real world, there is almost no absolute low rank. In addition, calculating the exact rank of a matrix is a morbid problem, because little changes of each element may lead to great changes in the rank of the matrix. [23]

According to PCA theory, [24] if matrix $X$ is low-rank, then the first $k(k \ll T)$ singular values occupy the majority of all singular values. [25]

In order to show the low rank of the traffic matrix on different working days in detail, we show the proportion of the first $k$ singular values of the traffic matrix $X$ collected over 5 working days, which is shown in Figure 3. The top 50 singular values occupy over 80% variance of the sum variance captured by all the singular values. In addition, the low rank of Tuesday, Wednesday, and Friday is better, which indicates that the traffic flow matrix has a good low rank property.

3.2 | Temporal stability property

Generally speaking, the difference between each pair of adjacent time slots monitored by a same traffic flow sensor node will not be too large, or even hardly changed. We call this characteristic temporal stability, in order to prove it, define the gap between each pair of adjacent time slots as

$$\Delta(t, n) = |x_{t,n} - x_{t-1,n}|$$ (2)

where, $t$ represents time and $n$ represents the detector location. Obviously, there are $2 \leq t \leq T$, $1 \leq n \leq N$, and the smaller $\Delta(t, n)$ indicates that the $n$-th sensor node collects data at time slot $t$ compared to time slot $t - 1$. In order to better reflect the overall characteristics of $\Delta(t, n)$, calculate the normalized gap

$$\bar{\Delta}(t, n) = \frac{|x_{t,n} - x_{t-1,n}|}{\max\{|x_{2,1} - x_{1,1}|, ..., |x_{T,n} - x_{T-1,n}|\}}$$ (3)

where $\max\{|x_{2,1} - x_{1,1}|, ..., |x_{T,n} - x_{T-1,n}|\}$ is the maximal gap between any two adjacent time slots of the $n$-th sensor, thus the smaller $\bar{\Delta}(t, n)$ indicates that $\Delta(t, n)$ is relatively smaller.
Similarly, we have analysed the temporal stability property of the traffic flow over five working days, and Figure 4 shows the Cumulative Distribution Function (CDF) of different $\Delta(t, n)$ from Monday to Friday, it is obvious that approximately 90% of the normalized adjacent gap is less than 0.4. Concern the proportion of equal adjacent data is more than 10%, which indicates that the gap of double adjacent traffic data collected by the same sensor is small, hence traffic flow matrix is temporal stable.

4 | THE PROPOSED METHODOLOGY

4.1 | RTLRR

Let $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{T \times N}$ be a complete traffic flow matrix, $x_i \in \mathbb{R}^T$ is a column vector, that is, the $i$-th sample, where $T$ represents quantity of the time slot, that is $T = 288$. In this article, in order to fully explore the value of traffic data, $N$ is the product of the number of flow sensors and the number of sampling days, therefore, the model we proposed takes not only the spatial correlation of different sensors into account, but also the temporal correlation of each sensor, so $N = 63 \times 20 = 1260$.

In representation theory, each sample in $X$ can be represented by a linear combination of samples in a dictionary $A = [a_1, a_2, \ldots, a_z] \in \mathbb{R}^{T \times z}$, that is,

$$X = AW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are

$$x_i = \sum_{j=1}^{z} a_j w_i(j)$$

where $w_i(j)$ represents the $j$-th element of $w_i$. However, dictionary is often overcomplete, therefore formula (5) is prone to achieve a trivial solution. A good solution is to replace $A$ with $X$. This is Self Representation theory, [21] then formula (4) is rewritten as follows

$$X = XW$$

Note that $X$ replaces $A$, and it is easy to know that $W \in \mathbb{R}^{N \times N}$.

In sparse representation theory, $W$ is a sparse matrix. However, sparse representation has the disadvantage of not being able to capture the global properties of $X$ which means it cannot make full use of the inherent information of traffic data in our problem, what is more, sparse representation is sensitive to noise. Therefore, Liu et al. [26] proposed LRR theory which is different from SR which solves the sparse representation of each sample independently. The goal of LRR is to jointly solve the representation of the lowest rank. $W$ is a low-rank matrix, so the problem can be modelled as a rank minimisation problem

$$\min_{W} \text{rank}(W) \quad \text{s.t.} \quad X = XW$$

where $W = [w_1, w_2, \ldots, w_N]$ is a coefficient matrix, and $w_i$ corresponds to $x_i$. Therefore, there are
However, the optimisation problem of formula (7) is an NP problem. Fortunately, Candès et al. [16] proved that the above rank minimisation problem can be relaxed as a nuclear norm minimisation problem

$$\min_{WW^T} \|WW\|_* \quad \text{s.t. } X = XW$$

where $\| \cdot \|_*$ is the nuclear function, which is the sum of non-zero singular values. Further, the above formula can be written in unconstrained form through penalty function [27] as

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_F$$

s.t. $\Omega = P_\Omega(M)$

(9)

where $\lambda_1$ is a hyperparameter.

Notice that the original $X$ is incomplete in fact, note the observed $X$ as $M \in \mathbb{R}^{T\times N}$, that is, the sampling matrix, then $P_\Omega(X) = P_\Omega(M)$, and $P_\Omega(\cdot)$ is an orthogonal projection operator, defined as

$$[P_\Omega(X)]_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega \\ 0, & \text{else} \end{cases}$$

(10)

where, $\Omega$ is the index of element observed in $X$. Thus problem (9) becomes

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_F$$

s.t. $P_\Omega(X) = P_\Omega(M)$

(11)

Note that there is a solution target $X$ in the projection operator, which is not conducive to subsequent solutions, so equation $P_\Omega(X) = P_\Omega(M)$ can be transformed into $M = X + E$, where $P_\Omega(E) = 0$, so (11) is equivalent to the following

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_*$$

s.t. $M = X + E$, $P_\Omega(E) = 0$

(12)

where $E$ is the missing error.

In Section 3.2, it was experimentally proved that the traffic flow data has temporal stability. Thus, we utilize this prior information via fused lasso, there are

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_1$$

s.t. $M = X + E$, $P_\Omega(E) = 0$

(13)

where $R \in \mathbb{R}^{(T-1)\times T}$ is a Toeplitz matrix

$$R = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -1 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

(14)

By minimizing the L1 norm regularisation term $\|RX\|_1$, the performance of temporarily stabilizing the elements in the normal traffic flow matrix in the time dimension can be obtained.

Furthermore, sampling data probably has noise, we note noise matrix as $C \in \mathbb{R}^{T\times N}$, to make our model robust, we add the L2 norm regularisation over $C$

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_1 + \frac{\lambda_3}{2}\|C\|_F^2$$

s.t. $M = X + C + E$, $P_\Omega(E) = 0$

(15)

In addition, the traffic flow data is non-negative, so $X \geq 0$ must be added. Therefore, by adding this constraint, problem (15) is further improved as

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_1 + \frac{\lambda_3}{2}\|C\|_F^2$$

s.t. $X \geq 0$, $M = X + C + E$, $P_\Omega(E) = 0$

(16)

### 4.2 Optimisation algorithm

For the detachable convex optimisation problem, the simplest and most efficient solution is the Alternating Direction Method of Multipliers (ADMM). ADMM converts the complex original problem into several easy-to-solve sub-problems. By seeking the local optimal solution of the sub-problems, the global optimal solution of the original problem can be found. We use ADMM proposed by Lin et al. [28] to solve the optimisation problem in this which do not need to adjust optimal penalty parameters.

To solve the optimisation problem of Equation (16), we first define the following two indicator functions

$$g(X) = \begin{cases} 0, & X \geq 0 \\ +\infty, & \text{else} \end{cases}$$

(17)

and

$$f(E) = \begin{cases} 0, & P_\Omega(E) = 0 \\ +\infty, & \text{else} \end{cases}$$

(18)

Then, the above optimisation problem can be transformed into the form by penalty function as Equation (19), where $\lambda_3$ is a penalty parameter for violating the constraint term.

Therefore, the optimisation process can be equivalently converted to solving the following three sub-problems iteratively of (20), where $X^k$, $W^k$ and $E^k$ respect their own $k$-th iteration result respectively.

$$\min_{X,W,E} \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|WW\|_* + \lambda_2\|RX\|_1 + g(X) + f(E) + \frac{\lambda_3}{2}\|M - X - E\|_F^2$$

(19)
\[
\begin{align*}
X^k &= \arg\min_{X} \frac{1}{2} \|X - X^{k-1}\|_F^2 + \lambda_3 \|M - X\|_F^2 + g(D) + \frac{\lambda_3}{2} \|M - X - E^{k-1}\|_F^2 \\
W^k &= \arg\min_{W} \lambda_1 \|W\|_* + \frac{1}{2} \|X^k - X^k W\|_F^2 \\
E^k &= \arg\min_{E} f(E) + \frac{\lambda_3}{2} \|E - (M - X^k)\|_F^2
\end{align*}
\]
(20)

\[
L(X, D, S, Y_1, Y_2) = \frac{1}{2} \|X - XW\|_F^2 + \lambda_3 \|M - X\|_F^2 + g(D) + \frac{\lambda_3}{2} \|M - X - E\|_F^2 + \langle Y_1, S - RX \rangle \\
+ \langle Y_2, D - X \rangle + \frac{\mu_1}{2} \left( \|S - RX\|_F^2 + \|D - X\|_F^2 \right)
\]
(21)

\[
X' = \arg\min_{X} \frac{1}{2} \|X - XW\|_F^2 + \lambda_3 \|M - X\|_F^2 + \frac{\mu_1}{2} \left( \|S - RX + H_{1}^{-1}\|_F^2 + \|D - X + H_{2}^{-1}\|_F^2 \right)
\]
(22)

\[
Q(X) = \frac{1}{2} \|X - XW\|_F^2 + \lambda_3 \|M - X\|_F^2 + \frac{\mu_1}{2} \left( \|S - RX + H_{1}^{-1}\|_F^2 + \|D - X + H_{2}^{-1}\|_F^2 \right)
\]
(23)

Obviously, the above sub-problems 2 and 3 have closed-form solutions, while sub-problem 1 is difficult to solve due to the existence of \(\|RX\|_1\) and \(g(X)\).

**For sub-problem 1:** In order to efficiently solve, we may introduce variables \(D\) and \(S\), let \(D = X\), \(S = RX\), then sub-problem 1 can be transformed into the following equivalent constraint optimisation problem

\[
\begin{align*}
\min_{X, D, S} & \frac{1}{2} \|X - XW\|_F^2 + \lambda_3 \|M - X\|_F^2 + g(D) + \lambda_3 \|M - X - E\|_F^2 \\
\text{s.t. } & S = RX, \ D = X
\end{align*}
\]
(24)

where, for simplicity, \(W^{k-1}\) is abbreviated as \(W\), \(E^{k-1}\) is abbreviated as \(E\).

We apply ADMM to solve the above sub-problems, the augmented Lagrange function can be defined as Equation (21), where \(Y_1\) and \(Y_2\) is Lagrange Multipliers, provided \(H_1 = Y_1 / \mu_1\) and \(H_2 = Y_2 / \mu_1\), the optimisation steps are

**Step 1:** Update \(X\): see Equation (22)

**Step 2:** Update \(D\):

\[
D' = \arg\min_{D} g(D) + \frac{\mu_1}{2} \|D - (X' - H_{2}^{-1})\|_F^2
\]
(25)

then we get

\[
\begin{align*}
D' &= X' - H_{2}^{-1} \\
D &\geq 0
\end{align*}
\]
(26)

**Step 3:** Update \(S\):

\[
S' = \arg\min_{S} \lambda_2 \|S\|_1 + \frac{\mu_1}{2} \|S - (RX' - H_{1}^{-1})\|_F^2
\]
(27)

this is the nearest neighbour operator of \(L1\) norm.

**Step 4:** Update \(H_1, H_2\):

\[
H_1' = H_{1}^{-1} + (S' - RX') \\
H_2' = H_{2}^{-1} + (D' - X')
\]
(28)
(29)

**Step 6:** Update \(\mu_1\):

\[
\mu_1 = \min(\rho \mu_{1-1}, \mu_{\max})
\]
(30)

The closed-form solution analysis of update \(X\) is as Equation (23) shown (for simplicity, the superscript of the number of iterations has been ignored), where, \(I\) is an identity matrix of \(\text{NXN}\).

Make

\[
\frac{\partial Q(X)}{\partial X} = 0
\]
(31)

\[
A = (I - W')(I - W')^T + (\lambda_3 I + \mu I + \mu RR^T)
\]
(32)

\[
B = \lambda_3 (M - E) + \mu (H_1 + S)R^T + \mu (D + H_2)
\]
(33)

then we have

\[
X = BA^{-1}
\]
(34)

**For sub-problem 2:** Assume that the SVD of \(X^k\) is \(X^k = U \Sigma V^T\), \(U = [U_1, U_2], \Sigma = \text{diag}(\Sigma_1, \Sigma_2)\), and \(\Sigma = [\Sigma_1^T \quad \Sigma_2^T]\), then the closed solution of sub-problem 2 is

\[
W^k = V_1 (I - \lambda_3 \Sigma_1^{-2}) V_1^T
\]
(35)

where \(I_1 = \{i: \alpha_i > \sqrt{\lambda_1}\}\) and \(I_2 = \{i: \alpha_i \leq \sqrt{\lambda_1}\}\).

**Proof:** In order to minimize \(\lambda_1 \|W\|_1 + \frac{1}{2} \|X^k - X^k W\|_F^2\), make the first order differential

\[
\lambda_1 \delta \|W\|_1 - \frac{1}{2} \|X^k W\|_F^2 = 0
\]
(36)
If \( W \)'s SVD is \( W = U_w \Sigma_w V_w^T \), then the sub-differential of \( W \)'s nuclear norm

\[
\partial \| W \|_W = \{ U_w V_w^T + Z, Z U_w^T Z = 0, Z V_w = 0, \| Z \|_2 \leq 1 \}
\]  

(37)

Let \( U_w = V_w = V_1 \), and then substitute (37) into (36), then

\[
\lambda_1 V_1 V_1^T + Z - (X^k)^T X^k(I - W) = 0
\]

(38)

Since

\[
I - W^k = I - V_1 \left( I - \lambda_1 \Sigma_1^{-2} \right) V_1^T = \lambda_1 V_1 \Sigma_1^{-2} V_1^T + V_2 \Sigma_2^{-2} V_2^T
\]

and

\[
(X^k)^T X^k = V \Sigma^2 V^T = V_1 \Sigma_1^{-2} V_1^T + V_2 \Sigma_2^{-2} V_2^T
\]

(39)

(40)

Substitute Equations (39) and (40) into Equation (38), we get

\[
\lambda_1 V_1 V_1^T + Z - \left( \frac{1}{\lambda_1} V_1 V_1^T + V_2 \Sigma_2^{-2} V_2^T \right) = 0
\]

(41)

then we have \( Z = \frac{1}{\lambda_1} V_1 \Sigma_1^{-2} V_1^T \), satisfy the conditions \( V_1^T Z = ZV_1 = 0 \) and \( \| Z \|_2 \leq 1 \).

For sub-problem 3: Its closed form solution is easily known as

\[
\begin{align*}
E^k &= M - X^k \\
P_\Omega(E^k) &= 0
\end{align*}
\]

(42)

So far, all the optimisation algorithms are proposed, the above solution process is named RTLRR-ADMM. It is worth mentioning that in order to speed up the optimisation speed of the algorithm, we skilfully updates \( X \) (this process is called inner loop), and then the obtained \( X \) is used to update \( W \) (this process is called outer loop), and the specific description is summarized in Algorithm 1.

## 5 EXPERIMENT AND ANALYSIS

### 5.1 Data preprocessing

Due to factors such as travel willingness, weather environment, and traffic accidents, traffic flow tends to show significant stochastic fluctuations or even surge, [29] such phenomenon is obvious in Figure 2. Data with the above characteristics can be regarded as an outlier. We regard the raw data as “corrupted” data which assuredly have noise and outliers, and use it to evaluate the estimation performance of each algorithm on true-world stochastic traffic dataset. In order to reduce the impact of such outliers on the traffic flow estimation performance, we preprocess the raw data by filtering. The original data and outlier filtered data within one day is shown in Figure 5. It can be seen that filtering process only significantly suppresses the phenomenon of fluctuations and surges, that is, outliers, which maintains stochastic nature of the raw data. In this article, raw data preprocessed by the filter is seen as “filtered” data.

![FIGURE 5 The raw data and filtered data](image-url)
5.2 Experiment configuration

For the purpose of comprehensively evaluating the model proposed in this article, we compare RTLRR with some well-known baseline methods, including MI, KNN, SVR, PPCA, ILRMD, and SSRp. Besides, in order to verify the role of the temporal regularisation term, the model without temporal regularisation is named RLRR. The parameters of several baseline methods are set according to the related work in [30], and the optimal parameters are determined by grid search.

The proposed algorithm and comparison algorithm were implemented in MATLAB 2016a, and the hardware environment is Intel (R) Core (TM) i5-8300H with 16GB RAM. In order to evaluate the estimation performance of each model on the missing value of traffic data, we used RStudio (package: http://www.stat.boogaart.de/compositions/) to design three patterns of missing data [22] as

(a) Missing Completely at Random (MCAR) Occurrence of data missing is completely random, independent of the existing data or other missing data, hence appears as isolated data points generally.

(b) Missing at Random (MAR) The occurrence of data missing is related to its neighbour points, which is generally reflected as continuous data points as a result.

(c) MIXED Half of the missing data is MCAR and the other half is MAR.

Figure 6 intuitively reflects the characteristics of these three patterns of missing data. Each column represents the traffic data sampled by one detector in one day and each row denote a time slot, the white squares represent the observed data points, and the black stand for the missing data points.

5.3 Evaluation criterion

The evaluation indicator used in this article is Root Mean Squared Error (RMSE), which depicts the absolute error of the estimation, and is defined as follows

\[
RMSE = \sqrt{\frac{1}{\text{Total}} \sum_{i=1}^{\text{Total}} (\hat{x}_i - x_i)^2}
\]  

where, Total is the number of missing data, \(x_i\) is the original data, and \(\hat{x}_i\) is the estimated data. Obviously, smaller RMSE indicates the better estimation performance.

In addition, in order to compare the estimation performances of each algorithm under different missing rate, we define the
missing ratio as $\delta$, then the total number of missing data is $\delta \times (N \times T)$.

### 5.4 Comparison of estimation performances

In this article, eight algorithms are executed 10 times on the same data set, and the average value of the results is used as the basis for evaluation. Tables 2–4 list the estimation performances of the above algorithms under three different missing patterns and different missing rates $\delta$. Where note the input sampling matrix $M$ without outliers as “filtered”, and $M$ contains noise and outliers as “corrupted”.

With respect to estimation performance, it can be clearly seen from the tables that each algorithm tends to be worse as the data missing rate increasing. It is reasonable that a higher data missing rate means much more serious information loss. What is interesting, MI’s estimation performance may not obey this law.

Secondly, no matter which algorithm is used, data outliers will cause a decrease in estimation performance. In particular, the sensitivity of each algorithm to noise varies greatly. For example, ILRMD and SSRp are very sensitive to noise, while RLRR and RTLRR are robust to noise which are based on LRR. It is particularly interesting that MI is also insensitive to noise, but MI’s estimation performance is the worst. What is more, MCAR missing pattern is the easiest situation to handle with, while MAR missing pattern is the toughest, MIXED missing pattern is centered. Also, RMSE of every model on corrupted dataset is higher than on filtered dataset, which indicates that outliers have a certain negative impact on estimation performance. Finally, no matter which type of data and missing type, the performance of MI among the eight algorithms is the worst, and the gap towards other algorithms is very obvious, then KNNR and SVR, PPCA and ILRMD have the similar performance, respectively. It is particularly noteworthy that as a state-of-art method, the performance of SSRp based on the sparse representation theory is close to RLRR based on the LRR theory, but is inferior to RTLRR, which indicates that utilizing the temporal correlation
TABLE 4  Estimation error (RMSE) under MIXED missing pattern

| 𝛿   | Data type | MI   | KNNR | SVR  | PPCA | ILRMD | SSRp | RLRR | RTLRR |
|-----|-----------|------|------|------|------|-------|------|------|-------|
| 0.1 | Filtered  | 223.12 | 114.53 | 103.36 | 95.75 | 88.23 | 82.25 | 75.35 | 67.89 |
|     | Corrupted | 225.40 | 118.27 | 109.55 | 102.12 | 93.62 | 87.38 | 78.43 | 70.47 |
| 0.2 | Filtered  | 227.76 | 119.19 | 113.27 | 97.37 | 91.57 | 88.29 | 79.21 | 70.32 |
|     | Corrupted | 225.24 | 125.40 | 119.16 | 105.79 | 96.39 | 94.37 | 83.07 | 74.17 |
| 0.3 | Filtered  | 220.45 | 126.38 | 116.54 | 104.62 | 94.28 | 93.51 | 84.62 | 74.29 |
|     | Corrupted | 225.53 | 132.17 | 123.29 | 111.61 | 100.13 | 99.34 | 89.55 | 79.18 |
| 0.4 | Filtered  | 228.22 | 131.87 | 125.35 | 107.12 | 99.32 | 99.05 | 87.10 | 75.30 |
|     | Corrupted | 236.46 | 138.05 | 133.54 | 115.22 | 106.60 | 107.44 | 93.23 | 80.04 |
| 0.5 | Filtered  | 226.69 | 137.79 | 141.26 | 111.90 | 111.24 | 108.80 | 89.05 | 78.31 |
|     | Corrupted | 230.95 | 148.06 | 149.70 | 119.37 | 119.16 | 116.29 | 96.18 | 84.81 |
| 0.6 | Filtered  | 229.64 | 145.20 | 151.32 | 116.46 | 118.37 | 114.08 | 92.33 | 82.73 |
|     | Corrupted | 234.23 | 152.66 | 158.51 | 125.55 | 125.43 | 122.61 | 99.33 | 89.42 |

TABLE 5  Run time (s) of different models under MIXED missing pattern

|     | MI    | KNN   | SVR   | PPCA  | ILRMD | SSRp | RLRR  | RTLRR |
|-----|-------|-------|-------|-------|-------|------|-------|-------|
| 0.1 | 1.578 | 18.431 | 22.620 | 30.190 | 38.018 | 31.282 | 33.229 |

FIGURE 7  Estimation error of RLRR and RTLRR

of traffic data via fused lasso can improve the estimation accuracy. As we can conclude, the proposed RTLRR achieves the best performance.

In order to compare the run time of different models, Table 5 shows different run time of the 8 models under MIXED missing pattern with the missing ratio 𝛿= 0.3. As we can see, MI performs fastest among all the methods for its simplicity, while methods which require iterative optimisation like SSRp, RLRR and RTLRR run slower obviously. Also, probabilistic based model PPCA requires more run time for its large amount of calculation. Benefit from ADMM optimisation algorithm, RTLRR we proposed runs faster than the similar model SSRp.

Considering the estimation error of different methods, we may draw the conclusion that RTLRR achieves significant improvement in estimation performance with moderate increase in run time.

For the purpose of visually comparing the performance of the presence or absence of a temporal regularisation term on the estimation performance, Figure 7 shows the estimation performance of RLRR and RTLRR on the clean data set and the corrupted data set under MIXED missing pattern. It can be seen from the figure that the RMSE of RTLRR is always lower than that of RLRR regardless of filtered or corrupted dataset, which verifies the significance of the addition of temporal regularisation term. In addition, it can be seen that the temporal regularisation term could improves the robustness of the model to a certain extent.

6  CONCLUSION

In this article, we propose a novel model RTLRR, which makes full use of two inherent prior information, low-rank property and temporal stability property of the traffic matrix, thus has achieved excellent results on real datasets. Different from the previous various modelling methods, we use the low-rank representation theory to model the spatial correlation of traffic data, furthermore, we model the temporal correlation via fused lasso. This method can effectively estimate the missing values in noisy traffic data due to the noise robustness of LRR. In order to solve this model, an optimisation algorithm based on ADMM is developed to solve alternately, which greatly accelerates the optimisation speed. Simulation results confirm the efficiency of the proposed method. Our future work will involve graph structure of road networks into modelling in order to further improve estimation performance. We also plan to combine other machine learning models with our proposed RTLRR to further reduce estimation error.
ACKNOWLEDGEMENTS
This work was supported in part by the National Natural Science Foundation of China under Grant 61812190, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20161104.

ORCID
Ruyong Mao https://orcid.org/0000-0002-5644-2305

REFERENCES
1. Li, D.M., et al.: Intelligent transportation system in macao based on deep self-coding learning. IEEE Transactions on Industrial Informatics 14, 3253–3260 (2018)
2. Zhou, W., et al.: A multi-source based coupled tensors completion algorithm for incomplete traffic data imputation. In: 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi’an, China, 23–25 October 2019
3. Chen, X., et al.: Spatiotemporal variable and parameter selection using sparse hybrid genetic algorithm for traffic flow forecasting. International Journal of Distributed Sensors Networks 13(6), 1–14 (2017)
4. Tian, D., et al.: Self-organized relay selection for cooperative transmission in vehicular ad-hoc networks. IEEE Trans. Veh. Technol. 66(10), 9534–9549 (2017)
5. Wu, Y., et al.: A hybrid deep learning based traffic flow prediction method and its understanding. Transportation Research Part C: Emerging Technologies 90, 166–180 (2018)
6. Feng, X., et al.: Adaptive multi-kernel SVM with spatial-temporal correlation for short-term traffic flow prediction. IEEE Trans. Intell. Transp. Syst. 20, 2001–2013 (2018)
7. Tak, S., Woo, S., Yeo, H.: Data-driven imputation method for traffic data in sectional units of road links. IEEE Trans. Intell. Transp. Syst. 17(6), 1762–1771 (2016)
8. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17(5), 519–533 (2003)
9. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. Roy. Statist. Soc. B 61(3), 611–622 (1999)
10. Cai, J., Cand’é, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20, 1956 (2010)
11. Robert, T., et al.: Sparsity and smoothness via the fused lasso. Journal of The Royal Statistical Society Series B-statistical Methodology 67(1), 91–108 (2005)
12. Shang, Q., Yang, Z., Gao, S., Tan, D.: An imputation method for missing traffic data based on FCM optimized by PSO-SVR. Journal of Advanced Transportation 2018, 2935248 (2018)
13. Wu, J., et al.: Bayesian optimization with gradients. In: Advances in Neural Information Processing Systems (NIPS), Long Beach, 4–9 December 2017
14. Lorenz, R., et al.: Dissociating frontoparietal brain networks with neuroadaptative Bayesian optimization. Nature Communications 9, 1227 (2018)
15. Xiao, F., et al.: Anomaly-tolerant network traffic estimation via noise-immune temporal matrix completion model. IEEE J. Sel. Areas Commun. 37(6), 1192–1204 (2019)
16. Cand’e, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. 9(6), 717–772 (2009)
17. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory 56, 2980–2998 (2010)
18. Dai, W., Milenkovic, O., Kerman, E.: Subspace evolution and transfer (SET) for low-rank matrix completion. IEEE Trans. Signal Process. 59, 3120–3132 (2011)
19. Luo, X., et al.: Traffic data imputation algorithm based on improved low-rank matrix decomposition. Journal of Sensors 2019, 1–11 (2019)
20. Chen, X., He, Z., Sun, L.: A bayesian tensor decomposition approach for spatiotemporal traffic data imputation. Transportation Research Part C: Emerging Technologies 98, 73–84 (2019)
21. Zhang, Q., et al.: Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. Inf. Fusion 40, 57–75 (2018)
22. Chen, X., et al.: Nonconvex lp-norm regularized sparse self-representation for traffic sensor data recovery. IEEE Access 6, 24279–24290 (2018)
23. Markovsky, I.: Low Rank Approximation: Algorithms, Implementation, Applications. Springer, (2011)
24. Choi, Y., Shon, H., Byon, Y., et al.: Enhanced application of principal component analysis in machine learning for imputation of missing traffic data. Application Science 9, 2149 (2019)
25. Xie, K., et al.: Low cost and high accuracy data gathering in WSNs with matrix completion. IEEE Transactions on Mobile Computing 17(7), 1595–1608 (2018)
26. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: International Conference on Machine Learning (ICML), Haifa, 21–24 June 2010
27. Liu, Q., Xu, Y., Zhou, Y.: A class of exact penalty functions and penalty algorithms for nonsmooth constrained optimization problems. Journal of Global Optimization 76, 745–768 (2020)
28. Lin, Z., Liu, R.S., Su, Z.X.: Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. Mach. Learn. 99(2), 287–325 (2015)
29. Carresse, S., et al.: Dynamic demand estimation and prediction for traffic urban networks adopting new data. Transportation Research Part C: Emerging Technologies 81, 83–98 (2017)
30. Fan, J., Chow, T: Matrix completion by least-square, low-rank, and sparse self-representations. Pattern Recognit. 71, 290–305 (2017)