Data Matrix Completion Based on Pattern Classification

Siyuan Lu
College of Information Engineering, Capital Normal University, Beijing, China
siyuanlu129@163.com

Xiaolan Tang*
College of Information Engineering, Capital Normal University, Beijing, China
tangxl@cnu.edu.cn

Yu Liu
College of Information Engineering, Capital Normal University, Beijing, China
liuyu@cnu.edu.cn

Wenlong Chen
College of Information Engineering, Capital Normal University, Beijing, China
chenwenlong@cnu.edu.com

ABSTRACT

In recent years, with the rapid development of big data technology, the matrix completion is often used for data recovery, and how to improve the accuracy of matrix completion is a key issue. This paper proposes a matrix completion method based on pattern classification, called PCRE, to improve data recovery performance. Since the hidden similarity within the data is a significant factor affecting the overall performance, the method PCRE uses non-negative matrix decomposition to extract the patterns of the data and accordingly rearranges the data matrix to fit for the matrix completion. Experiments are conducted by using PM 10 monitoring data collected by 34 sensors in Beijing in 2019 (totally 351 days). The results show that, compared with existing methods, PCRE improves the accuracy of data recovery with a shorter computation time.

CCS CONCEPTS

• **Applied computing**; • **Computer forensics**; • **Data recovery**;

KEYWORDS

Data recovery, Matrix completion, Matrix decomposition, Pattern classification

ACM Reference Format:
Siyuan Lu, Xiaolan Tang*, Yu Liu, and Wenlong Chen. 2021. Data Matrix Completion Based on Pattern Classification. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487091

1 INTRODUCTION

In the information era, massive data are collected by different kinds of devices and analyzed by various data processing algorithms. Due to some reasons, such as a high sampling cost or an equipment failure, sometimes only a part of the data is gathered. In other words, some data is lost or invalid. When the collected data is incomplete, the further use of data may not obtain desired results. For example, the sensors monitoring the air quality may lose the data at a certain time due to the device failure, and it may lead to wrong data about air quality at this time and hence affect the people's health. Therefore, the recovery of missing data is important for the social production and the people's lives.

As one of the main approaches of data recovery, matrix completion effectively fills the missing items according to some known information, so as to obtain more complete and accurate information for users to make decision. It has been found that the matrix completion performance is closely related to the similarity inside the data [1]. However, the data correlation has not been fully utilized in the process of data recovery. How to improve the matrix completion considering the similar features of data is a key problem.

This paper proposes a matrix completion method based on pattern classification, called PCRE. The method uses a non-negative matrix decomposition algorithm to extract the patterns of the data, and then the data are sorted and rearranged according to the patterns. In this way, the similar data are put together as a block, and hence the matrix completion method taking the rearranged data as input data has a higher recovery accuracy.

The main contributions of this paper are listed below.

- A matrix completion algorithm based on pattern classification is proposed, which consists of two main steps. Firstly, a pattern classification method based on non-negative matrix decomposition is proposed to fully explore the data similarity. Then a matrix rearrangement method based on the patterns is designed to reduce the lower sampling limit.
- We use Beijing PM 10 data in 2019[2] as an example to test the recovery performance of the matrix completion method based on pattern classification. We define a matrix $M_{N\times T}$ to present PM 10 data, in which N is the number of sensors and T is the number of time slots. A popular matrix completion algorithm OptSpace is selected, and the OptSpace and PCRE-OptSpace are compared. The experiments demonstrate the accuracy and the time-efficiency of the method proposed in this paper.

The full paper is organized as follows.

- Section 1 introduces the research background, the brief introduction to our work and the main contributions of this paper.

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487091

- Section 2 introduces the related work to relevant techniques in this paper and provides an introduction to the common algorithms for matrix completion and feature extraction.
- Section 3 proposes a pattern classification method, uses non-negative matrix decomposition to mine the implied relevance of the data, and then rearranges data matrix according to the patterns.
- Section 4 presents the performance evaluation, in which we compare the original matrix completion algorithm OptSpace and the matrix completion integrated with our pattern classification and data rearrangement, called PCRE-OptSpace.
- Section 5 concludes the paper and provides an outlook for future work.

2 RELATED WORK

Candés et al [3] proved that a low-rank matrix with size $n_1 \times n_2$ and rank r can be recovered by solving a simple convex optimization problem when a sufficient number of samples are provided.

The article demonstrates that the sampling rate needs to satisfy the condition m $\geq Cn^{6/5} r \log n$, where C is a constant, and n $= \max\{n_1, n_2\}$. Both r and n affect the value of m, and the lower limit on the number of samples required for matrix recovery can be reduced by reshaping the matrix shape, which in turn leads to better recovery performance [4].

An inadequate number of samples usually lead to a long time to recover the missing data and a low accuracy of the recovered data, sometimes the matrix completion algorithm may not converge. Moreover, although matrix completion performs better than other approaches for the data with low sampling rate, its performance is greatly affected when the missing rate is very high. Therefore, the matrix completion aiming at accurate data recovery at low sampling rate attracts more attentions.

Because the correlation property within the data causes the data to be sparse, the sparsity of data makes it a feasible way to infer the rest from the collected part. Qu et al.[5] demonstrated that similarity is an important factor affecting the recovery performance of matrix completion.

Based on the analysis of a large amount of air quality monitoring data, Wang et al.[6] exposed the potential temporal stability, spatial correlation and other characteristics of air quality monitoring data. Then they proposed a matrix rearrangement principle, which can reduce the lower sampling limit.

Peng et al.[7] [8]used the non-negative matrix decomposition method [9] to mine the internal similarity characteristics of the data, and processed the data accordingly, which effectively improve the utilization of data similarity.

At present, the matrix completion algorithms mainly include four categories, small-scale matrix completion algorithms, kernel parametric minimization algorithms, Grassmannian manifold minimization algorithms, and other novel algorithms [10] [11].

The Grassmannian manifold minimization solutions including the OptSpace algorithm [17], the SET algorithm [1]and others. Other novel algorithms include low-rank matrix fitting algorithm [18], truncated kernel parametric algorithm [19]. Chen et al.[20]also develop a procedure to compensate for the bias of the widely used convex and non-convex estimators. The resulting de-biased estimators admit nearly accurate non-asymptotic distributional guarantees.

The existing matrix completion algorithms have some limitations. Firstly, the data recovery performance will be greatly affected if the data missing ratio is high. Secondly, as the rank or dimension of the data matrix increases, the computation time increases significantly and the relative error ratio of matrix recovery will be larger [3] [10] [21].

In order to improve the recovery performance of matrix completion, in this paper we explore the similarity properties implied within the meteorological data PM 10, and rearrange the data accordingly to reduce the lower sampling limit and increase the accuracy of data recovery.

3 PATTERN CLASSIFICATION AND MATRIX REARRANGEMENT

Existing research [6] showed that meteorological data implies similar features including the temporal periodicity, the location correlation, etc. In this section, we explore the hidden patterns of the monitoring data by using the non-negative matrix decomposition, and then rearrange the data matrix according to the similarity. We use 2019 Beijing PM 10 monitoring data as the original data to evaluate our data recovery method based on pattern classification. The dataset is obtained from Beijing Municipal Ecological and Environmental Monitoring Center [2]. The data were collected by 34 sensors widely distributed in Beijing.

We define the matrix $M_{N\times T}$, in which N is the number of sensors and T is the number of time slots. m_{nt} is the data measured by the nth sensor at the tth time slot.

3.1 Pattern Classification

3.1.1 Pattern Quantity. When performing a non-negative matrix decomposition, the number of fundamental patterns q needs to be determined. We select the proper value of q according to the background features of the specific problem. In this scenario, we assume the initial range of q from 3 to 14, and we use each value of q to conduct the non-negative matrix decomposition. Assuming that there are n data matrices (a data matrix is the data collected during a day), there are n decomposition results for each value of parameter q. Calculate the Euclidean distance D for each row of these matrices, and then the value of q having the smallest D is selected.

3.1.2 Matrix Decomposition. After determining q, the coefficient matrix C and the pattern matrix P are obtained by performing non-negative matrix decomposition (the number of basic patterns is set to q). The non-negative decomposition of a matrix M is

$$\mathrm{M} \approx \mathrm{CP} \tag{1}$$

where C is the coefficient matrix, $C \in R_{+}^{N\times q}$, and P is pattern matrix, $P \in R_{+}^{q\times T}$.

Define the objective function as

$$\begin{aligned} &\min \sum_{\boldsymbol{i},\boldsymbol{j}} \left[\boldsymbol{M}_{\boldsymbol{ij}} - (\boldsymbol{CP})_{\boldsymbol{ij}}\right]^2 \\ &\text{s.t. } \; C \geq 0, P \geq 0 \end{aligned} \tag{2}$$

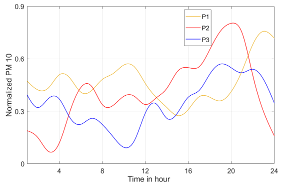


Figure 1: Patterns of Beijing PM 10.

It is used to get comparison of approximation degree between the matrix after decomposition and the original matrix.

To make the pattern matrix more accurate, we utilize the normalization method to minimize the difference with the following formula:

$$\begin{aligned} C_{ik} &\leftarrow C_{ik} \cdot \frac{(VP^T)_{ik}}{(CPP^T)_{ik}} \\ P_{kj} &\leftarrow P_{kj} \cdot \frac{(C^TV)_{kj}}{(C^TCP)_{kj}} \end{aligned} \tag{3}$$

3.1.3 Pattern Classification. Matrix C includes the coefficients of each sensor position with respect to the corresponding basic patterns. In other words, one row of C indicates the proportion and scale of each basic patterns at a sensor position. According to the mean value of C, we divide all sensors into q patterns. The pattern matrix P has some specific practical meanings. Since PM 10 is influenced by various types of human production, such as factory emissions, vehicle emissions, commercial area emissions, etc., the different patterns indicate different kinds of locations. 1 shows the three patterns of PM 10 over the day.

- Pattern 1 has a highest peak during 8am to 12am, and the overall concentration of PM 10 in pattern 1 is high. This is in line with city traffic flow. Therefore, pattern 1 can be regarded as a pattern deployed near the roads.
- Pattern 2 has peaks in the early hours and at night with significantly higher values than the other two patterns. It is similar to the change rules of exhausting emissions by factories. Thus, we regard pattern 2 as sensors near factories.
- Pattern 3 has peaks around morning, noon and night, which are basically consistent with the commute and meal times. Therefore, pattern 3 is regarded as a pattern near the commercial area.

In original PM 10 data matrix, the sensor data having the same pattern are not listed next to each other. In order to make full use of the implicit correlation and improve the completion performance, the sensor data in the original PM 10 data matrix is classified into 3 groups according to their patterns.

As shown in 2, in the original data, sensors with the same pattern are not adjacent; after obtaining the pattern information, we arrange the sensors having the same pattern to adjacent positions.

3.2 Matrix Rearrangement

Existing studies show that, data rearrangement (changing the shape of the data matrix) helps to improve the performance of matrix completion. Hence, in this section, we design a data rearrangement algorithm to adjust the data. In the original data matrix, the number of rows and the number of columns may be different, and sometimes the difference is huge. Existing research shows that if the matrix is rearranged into a square shape that has the same number of rows and columns, the matrix completion performance would be better [6]. Some studies proved that the segmentation of rows or columns based on period will not increase the rank of the matrix too much, and the dimensions of row space and column space of the rearranged matrix will not change. Therefore, the matrix is pruned according to the period h (which is usually determined by period of data change). If the number of rows is insufficient after clipping, some empty rows are added as a padding matrix, so that the rearranged data matrix has a square shape.

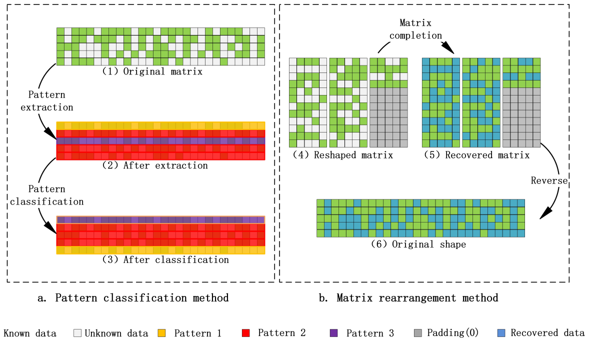


Figure 2: An example of PCRE.

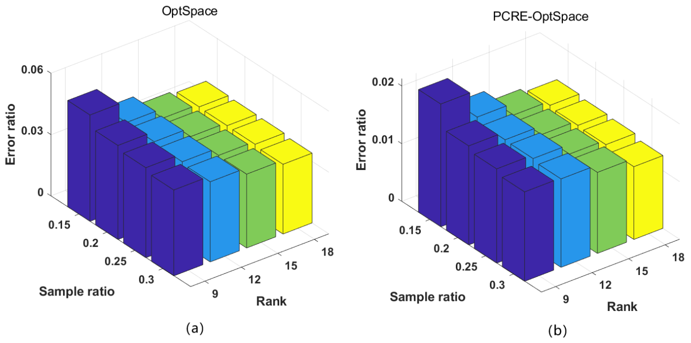
Algorithm 1 Matrix rearrangement algorithm

```
input: $M_{T\times N}$ $(T \gg N)$ , h;
output: W;
calculate the possible number of rows after rearrangement $p = \sqrt{T \times N}$;
calculate the number of rows in a sub-matrix m = $\frac{p}{h} \times h$;
calculate the number of sub-matrices that the original matrix M can be partitioned into $num = \frac{T}{\mathrm{m}}$ ;
split the original matrix M into $num$ sub-matrices;
if the number of rows in the last submatrix, l, is smaller than m, then
    add a padding matrix with m-l rows;
    fill the padding matrix with value 0;
end if
establish the reshaped matrix W by integrating $num$ sub-matrices;
return W;
```

Take 2 as an example. 2 **b** shows the process of matrix rearrangement and matrix completion. 2 **b (4)** shows the matrix rearranged according to the **Algorithm 1**, that has a square shape and the supplement rows are filled with zero. Then the matrix completion algorithm is used to recovery the rearranged matrix. Finally, the original shape of the matrix is reversed by conducting the inverse operations of the rearrangement algorithm.

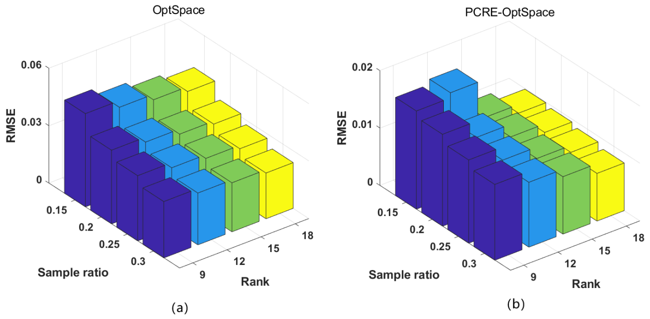
4 PERFORMANCE EVALUATIONS

In our experiments, the classical matrix completion algorithm OptSpace is used for data recovery, while the OptSpace with our



(a) (b)

Figure 3: Error Ratio Results.



(a) (b)

Figure 4: RMSE Results.

pattern classification and matrix rearrangement is called PCRE-OptSpace. OptSpace, as a Grassmannian flow minimization algorithm, is suitable for large-scale data processing and designed based on singular value decomposition algorithm to solve low rank matrix completion problem.

The experiments are carried out using the same computer server which is equipped with Intel(R) Core(TM) i7-5500U CPU @2.40 GHz and 8 GB memory in Windows 8.1 x64 OS.

In order to evaluate the data recovery performance, three metrics are analyzed, including error ratio (ER), root mean square error (RMSE) and calculation time (Time). In order to analyze the influence of the low sample ratio, the data sample ratio is set from 0.15 to 0.3. We change the rank of matrix from 9 to 18.

4.1 Error Ratio

The error ratio is calculated by

$$\mathrm{ER} = \frac{\sqrt{\sum_{(i,j)\in\bar{\Omega}} \left(w_{ij} - \hat{w}_{ij}\right)^2}}{\sum_{(i,j)\in\bar{\Omega}} \hat{w}_{ij}^2} \tag{4}$$

where w_{ij} and $\hat{w}_{ij}$ are the original and recovered data, respectively, and $\bar{\Omega}$ is a set containing the time slots and sensor locations with unknown data.

Note that only the unobserved data rather than the known data are used to compute the error ratio. A smaller error ratio indicates a smaller gap between the recovered data and the original data.

As shown in 3, compared with the original matrix completion solution OptSpace, the solution presented in this paper PCRE-based OptSpace (PCRE-OptSpace) achieves a much lower error ratio. Specifically, the mean error ratio of OptSpace is 0.039475000000000, while the mean error ratio of PCRE-OptSpace is 0.015236979166667. The PCRE method reduces the error ratio to 0.3860 times the original one. These experiments show that the PCRE scheme can improve the performance of Error ratio.

4.2 RMSE

Another metric RMSE is calculated by Unlike error ratio, in RMSE all data elements are involved in the calculation.

As illustrated in 4, PCRE-OptSpace achieves a much lower RMSE than OptSpace. The mean RMSE of OptSpace is 0.033999399038462, while the mean RMSE of PCRE-OptSpace is 0.012175000066600000, and the PCRE method reduces the RMSE to 0.3581 times the original one. This part shows that the PCRE scheme can improve the performance of RMSE.

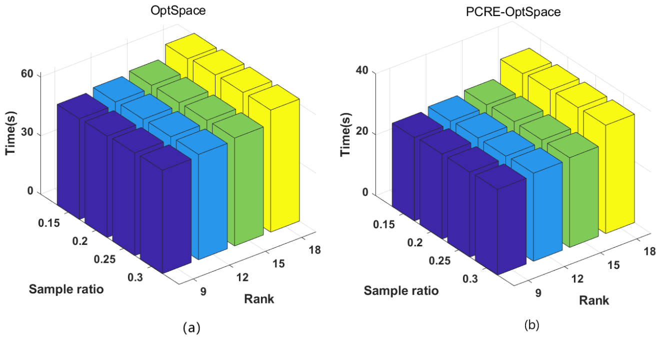
4.3 Time

We use the tic and toc functions to calculate the computation time.

In 5, our scheme PCRE-OptSpace achieves a much shorter calculation time. The mean time of OptSpace is 55.700000000000000, while the mean time of PCRE-OptSpace is 30.206477272727270. PCRE method reduces the calculation time to 0.5423 times the original one. This part shows that the PCRE scheme can cut down the computation time of matrix completion.

4.4 Experiment Conclusion

PCRE method is evaluated in several sets of experiments based on real Beijing PM 10 data, applying three performance metrics including error ratio, RMSE, and computation time. These experimental results above show that under low sample ratio, original matrix completion method (OptSpace) has low recovery accuracy and long computation time. Under the same conditions, PCRE improves data recovery performance by mining data similarity and lowering the lower sampling limit. The experiments demonstrate the effectiveness of the PCRE method.



(a) (b)

Figure 5: Computation Time Results.

5 CONCLUSION

In order to improve the accuracy and efficiency of data matrix completion, this paper proposes a matrix completion method based on pattern classification, named PCRE. Before the data completion, the data patterns are extracted from the original data by using the pattern analysis method, so as to classify the original data and efficiently discover the similarity of the data. Then, we rearrange the data matrix to move the similar data together and fill supplement matrix with zero. Experiments based on real PM 10 data in Beijing show that PCRE improves the accuracy of data recovery and reduces the computation time of data completion.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from the Beijing Natural Science Foundation (4202012), the National Natural Science Foundation of China (61872252), and the Science & Technology Project of Beijing Municipal Commission of Education in China (KM201810028017).

REFERENCES

[1] W.DAI,O.MILENKOVIC,E. KERMAN (2011). Subspace Evolution and Transfer form Low-Rank Matrix Completion[J].IEEE Transactions on Signal Processing, 59(7):3120-3132.

[2] PM 10 data in beijing, http://www.bjmemc.com.cn/jsps/jsp/zxgk/hjzljc.jsp.

[3] E. J. Candes, B. Recht (2008). Exact low-rank matrix completion via convex optimization, in: 2008 46th Annual Allerton Conference on Communication, Control, and Computing, pp. 806–812. doi:10.1109/ALLERTON.2008.4797640.

[4] E. J. Candes, T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion, IEEE Transactions on Information Theory 56 (5) 2053–2080. doi:10.1109/TIT.2010.2044061.

[5] Qu L, Li L, Zhang Y , *et al.* (2009). PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach[J]. IEEE Transactions on Intelligent Transportation Systems, 10(3):512-522.

[6] Le W (2017). Data collection and data completion based on the theory of sparse representation.

[7] Peng C, Jin X, Wong, K C, Shi, M., P Liò (2012). Collective human mobility pattern from taxi trips in urban area, PLOS ONE 7. doi:10.1371/journal.pone.0034487.

[8] Liu X, Liu, X, Wang Y, Pu J, Zhang X (2016). Detecting anomaly in traffic flow from road similarity analysis, Springer International Publishing. doi:10.1007/978-3-319-39958-4_8.

[9] D. D. Lee, H. S. Seung (1999). Learning the parts of objects by nonnegative matrix factorization, Nature 401 (7). doi:10.1038/44565.

[10] J. Deng, K. Xie (2017). Typical algorithm for low rank matrix completion, Electronic production (9) 40–43.doi:10.16589.

[11] Chen L, Chen SC (2017). Survey on matrix completion models and algorithms. RuanJianXueBao/Journal of Software, 28(6):1547−1564 (in Chinese). http://www.jos.org.cn/1000-9825/5260.htm

[12] Liu Z, L. Vandenberghe L (2009). Interior-Point Method for Nuclear Norm Approximation with Application to System Identification[J]. SIAM Journal on Matrix Analysis and Applications, 31(3):1235-1256.

[13] TOH K C, YUN S (2010). An Accelerated Proximal Gradient Algorithm form Nuclear Norm Regularized Least Squares Problems[J].Pacific J. Optimaization, 6.615-640.

[14] RECHT B, FAZEL M,PARRILO P A (2010). Guaranteed Minimum-Rank Solutions of Linear Matrix Equation Equations Via Nuclear Norm Minimization[J]. SIAM Rev., 52:471-501.

[15] CAI J, CANDES E J, SHEN Z (2010). A Singular Value Thresholding Algorithm for Matrix Completion[J] SIAM J. Op, 20:1956-1982.

[16] MAS, GOLDFARB D, CHEN L (2011). Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization[J]. Mathematical Programming, 12:321-353.

[17] Raghunandan H. Keshava, Andrea Montanari, Sewoong Oh (2010). Matrix completion from a few entries[J]. IEEE Transactions on Information Theory, 56(6):2980-2998.

[18] WEN Z,YIN W,ZHANG Y (2012). Solving A Low-Rank Factorization Model for Matrix Completion by A Nonlinear Successive Over-Relaxation Algorithm[J]. Math.Prog. Comp., 4(4):333-361.

[19] D. Zhang, Y. Hu, J. Ye, X. Li and X. He (2012). Matrix completion by Truncated Nuclear Norm Regularization," 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2192-2199.

[20] Chen Y, Fan J, Ma C, *et al.* (2019). Inference and uncertainty quantification for noisy matrix completion[J]. Proceedings of the National Academy of Sciences, 116(46): 22931-22937.

[21] Nguyen L T, Kim J, Shim B (2019). Low-rank matrix completion: A contemporary survey[J]. IEEE Access, 7: 94215-94237.