

Electricity Consumption Data Quality Improvement Based on Low-Rank Matrix Decomposition

Guo Xu, Xuyang Zeng, Tianyu Ma, Jianjun Xu *

School of Electrical Engineering and Information, Northeast Petroleum University, Daqing, Heilongjiang, 163318, China

* Corresponding author: Jianjun Xu (Email: xujj@nepu.edu.cn)

Abstract: Measurement data from smart meters can be affected by issues such as missing values and noise due to communication problems, power fluctuations, or interference. These anomalies reduce the reliability and utility of electricity consumption data, thereby hindering the efficient operation of power systems. To address the issues of missing data and noisy anomalies in electricity consumption data, this paper proposes a low-rank matrix recovery model based on the truncated nuclear norm. By leveraging the low-rank property inherent in electricity consumption data, the truncated nuclear norm is used as an approximation of the matrix rank. Simultaneously, the L1- norm is employed to constrain the sparse noise, enabling the completion of the low-rank components and the separation of sparse components. Finally, a two-step iterative strategy based on the alternating direction method of multipliers (ADMM) is applied to solve the model. Comparative experiments conducted on real datasets against commonly used methods for missing data imputation demonstrate the effectiveness of the proposed approach.

Keywords: Low-Rank Matrix Decomposition; Electricity Consumption Data Quality Improvement; Truncated Nuclear Norm; L1- norm; Alternating Direction Method of Multipliers.

1. Introduction

In recent years, power quality issues in low-voltage distribution networks have attracted widespread attention. With the extensive deployment of Advanced Metering Infrastructure (AMI), electricity information collection systems have achieved full coverage in low-voltage distribution networks. The application of big data technologies enables power companies to perform in-depth analysis of electricity consumption data. High-quality electricity data can significantly enhance the scientific rigor and accuracy of decision-making processes. However, due to various interferences affecting smart meters during operation, the collected data often exhibit anomalies, such as missing values and noise. These anomalies can adversely impact the effectiveness of data-driven applications. Therefore, addressing anomalous data is of critical importance for improving the accuracy of data mining processes.

In the field of data imputation, numerous studies have been conducted internationally. traditional interpolation methods are typically based on simple statistical models, which struggle to capture the complex nonlinear relationships among features [1]. With the advancement of artificial intelligence, neural networks have demonstrated the capability to learn high-dimensional, complex nonlinear relationships among input features, thereby offering more accurate imputation results [2]. However, neural networks require large amounts of training data to effectively learn the data distribution. If the sample size is insufficient, the network may fail to train effectively. Additionally, neural networks often require considerable training time, and this computational overhead increases significantly as the network size grows. Low-rank matrix completion leverages the assumption of matrix low-rank properties to impute missing data by identifying underlying structures. In existing studies on low-rank matrix recovery models, the nuclear norm is commonly used as an approximation of the matrix rank. However, the nuclear norm penalizes all singular values

equally during the iterative process. This means it not only penalizes singular values corresponding to noise but also unnecessarily penalizes those corresponding to the low-rank components. As a result, the nuclear norm is not the optimal surrogate for matrix rank.

2. Analysis of AMI Measurement Data Characteristics

In the actual operation of power system, due to the complexity of AMI working environment, the measurement data often show missing values and noise. The common deletion patterns are: random deletion and continuous deletion. Sparse noise is the most common type of noise.

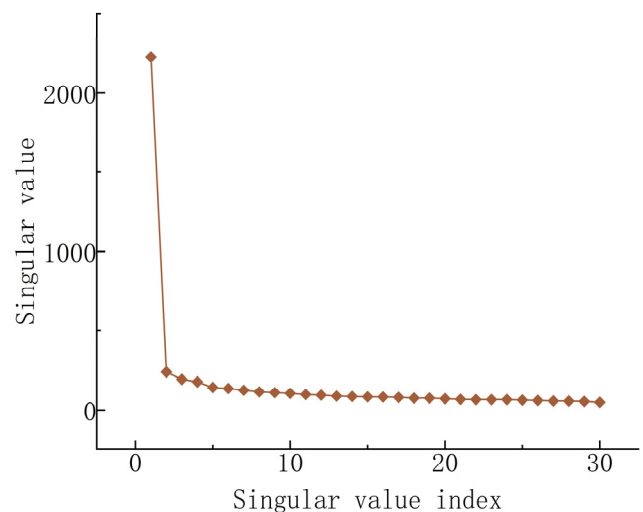


Figure 1. The singular values of the real electricity consumption data matrix arranged in descending order

The influence of various factors, such as the general usage patterns of electrical appliances, the guidance of user behavior through electricity pricing policies, and the power supply characteristics within a region, leads to a certain

degree of similarity in the electricity consumption behaviors of different households within the same area. Consequently, the overall electricity consumption data matrix exhibits an approximate low-rank characteristic.

The approximate low-rank property of a matrix can be evaluated through the singular value distribution [4]. Singular value decomposition was performed on the real electricity consumption data of low-voltage substations, as shown in Figure 1.

As shown in Figure 1, the singular value distribution of the electricity consumption data matrix exhibits significant unevenness, with the singular values rapidly decreasing and stabilizing at lower values. This suggests that the electricity consumption data of the low-voltage substations possesses certain low-rank characteristics.

3. The Construction and Solution of The Low-Rank Matrix Recovery Model Based on Truncated Nuclear Norm:

3.1. Model Construction

Based on low-rank space decomposition, perform decomposition on the observed matrix M :

$$M = X + E \quad (1)$$

Where: X is the low-rank component; E is the sparse noise component.

- The sparse noise is processed through the L_1 norm.
- The low-rank component X is processed through the rank function $rank(X)$.

Due to the discontinuous and non-convex nature of the rank function, it is typically approximated. Fazel proved that the nuclear norm is the tightest convex lower bound of the rank function of matrices. Therefore, the nuclear norm is generally used as an approximation of the rank function [5]. The nuclear

norm compresses all singular values during the iterative process, which may lead to the over-penalization of the low-rank components. In this paper, the truncated nuclear norm is chosen as a better approximation of the rank function. The truncated nuclear norm of a matrix A is defined as:

$$\|A\|_r = \sum_{i=r+1}^{\min(m,n)} \sigma_i(A) \quad (2)$$

The low-rank matrix model based on the truncated nuclear norm is formulated as follows:

$$\begin{aligned} \min & \|X\|_* - \max_{AA^T=I_{r \times r}, BB^T=I_{r \times r}} \text{Tr}(AXB^T) + \lambda \|E\|_1 \\ \text{s.t. } & P_\Omega(M) = X + E \end{aligned} \quad (3)$$

3.2. Model Solution:

For the optimization problem corresponding to equation (3), this paper solves it using a two-step iterative method:

- Given X_l , obtain A_l and B_l through eigenvalue decomposition:

$$X_l : [U_l, \Sigma_l, V_l] = \text{SVD}(X_l) \quad (4)$$

Where $U_l = (u_1, u_2, \dots, u_m) \in R^{m \times m}$,
 $V_l = (v_1, v_2, \dots, v_n) \in R^{n \times n}$; $A_l = (u_1, u_2, \dots, u_r)^T$,
 $B_l = (v_1, v_2, \dots, v_r)^T$;

- Given A_l and B_l update X_{l+1} and E_{l+1} :

$$\begin{aligned} \min & \|X\|_* - \max_{AA^T=I_{r \times r}, BB^T=I_{r \times r}} \text{Tr}(AXB^T) + \lambda \|E\|_1 \\ \text{s.t. } & P_\Omega(M) = X + E \end{aligned} \quad (5)$$

Select and update until $\|(X_{l+1} + E_{l+1}) - (X_l + E_l)\|_F \leq \varepsilon_0$,
Return the final results X, E .

Using the augmented Lagrangian method, equation (5) is transformed into an unconstrained optimization problem. The steps for solving are as follows.

$$\mathcal{L}(X, E, Y, \mu) = \|X\|_* - \text{Tr}(AXB^T) + \lambda \|E\|_1 + \langle Y, M - X - E \rangle + \frac{\mu}{2} \|M - X - E\|_F^2 \quad (6)$$

Step 1: update X :

$$\begin{aligned} X_{k+1} &= \arg \min \|X\|_* + \frac{\mu_k}{2} \left\{ \left\| X + \frac{H^T}{\mu_k} \right\|_F^2 - \left\| \frac{H^T}{\mu_k} \right\|_F^2 \right\} \\ H &= E_k - M - A_l B_l^T - Y_k^T \end{aligned} \quad (7)$$

Its closed-form solution is:

$$X_{k+1} = D_{2/\mu_k} \left(M - E + \frac{A_l^T B_l + Y_k}{\mu_k} \right) \quad (8)$$

Step 2: update E :

$$E_{k+1} = \arg \min_E \lambda \|E\|_1 + \frac{\mu}{2} \left\{ \left\| M + \frac{Y_k}{\mu_k} - X_{k+1} \right\|_F^2 \right\} \quad (9)$$

Its closed-form solution is:

$$E_{k+1} = S_{2\lambda/\mu} \left[M + \frac{Y_k}{\mu_k} - X_{k+1} \right] \quad (10)$$

Step 3: update Lagrange multiplier Y :

$$Y_{k+1} = Y_k + \mu [M - X_{k+1} - E_{k+1}] \quad (11)$$

Step 4: update Penalty factor μ :

$$\mu_{k+1} = \rho \mu_k, \rho > 1 \quad (12)$$

4. Low-voltage Substation Measurement Data Quality Improvement Tests

This section uses the Irish Social Science Data Archive dataset for verification. Piecewise linear interpolation, ADMM based on nuclear norm, and SVT are selected as comparison methods for this paper. Two missing patterns,

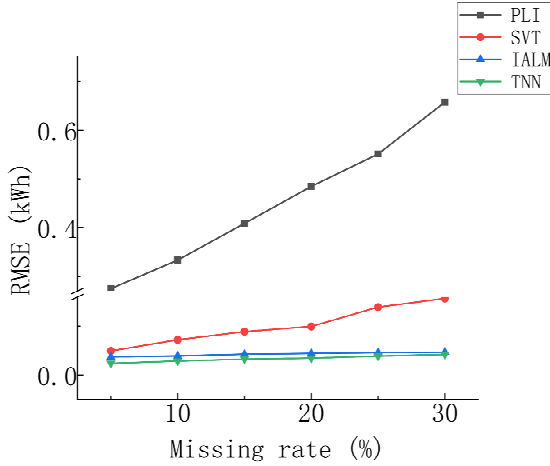
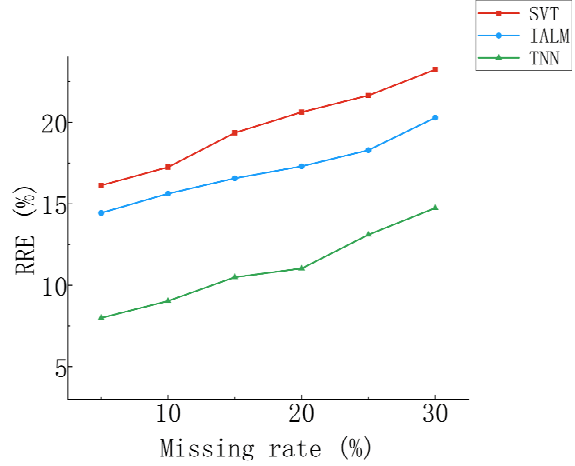


Figure 2. Comparison of data recovery effects under different random missing rates

continuous long-term missing and random missing, are considered.

First, the random missing pattern is verified. As shown in Figure 2, with the increase in the data random missing rate, the error significantly increases. However, the data recovery error of the proposed method is still significantly lower than that of the comparison methods. With a 20% missing rate, the RMSE and RRE of TNN are reduced by 21.76% and 19.81% compared to IALM, respectively.



Verify the continuous long-term missing pattern. Set that 20% of the natural days within the total time window have missing data, and within the natural days with missing data,

select different continuous time periods with missing data sequentially.

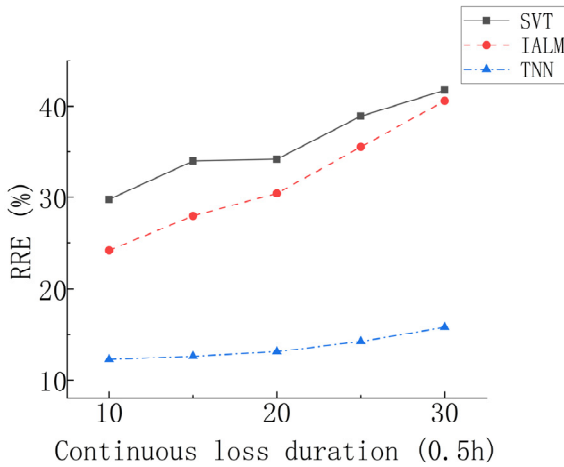


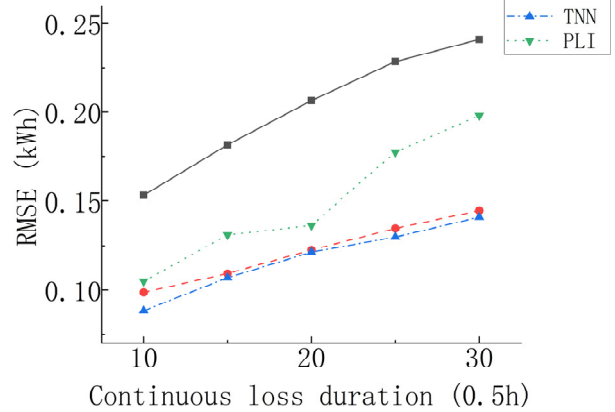
Figure 3. Comparison of data recovery effects under different continuous missing time periods.

As seen in Figure 3, under the same continuous missing ratio, the TNN recovery error outperforms other recovery methods. In the scenario with 20 continuous missing time periods, the RRE of TNN is only 13.11%, and the RMSE is only 0.121 kWh.

The experimental results indicate that, compared to common methods, the data recovery algorithm proposed in this paper achieves lower recovery errors when facing different data missing scenarios.

5. Conclusion

The paper proposes a data recovery algorithm based on truncated nuclear norm to address common anomaly issues in electricity consumption data. It uses truncated nuclear norm as an approximation of the matrix rank function and introduces norm regularization to achieve sparse noise



separation. For the non-convex model based on truncated nuclear norm, a two-step iterative algorithm based on ADMM is proposed. The algorithm is tested on publicly available datasets and compared with several commonly used algorithms, demonstrating that the proposed strategy provides good recovery results for various data missing scenarios. In future research, the study will further explore efficient strategies for selecting the r value of the truncated nuclear norm and the regularization coefficient λ .

References

- [1] Enders C K. Applied missing data analysis[M]. New York, USA: The Guilford Press, 2010:22-50.
- [2] Li X P, Wang M, So H C. An interpretable bi-branch neural network for matrix completion [J]. Signal Processing, 2022, 200: 108640.

- [3] Liao, Mang, et al. "An alternating direction method of multipliers based approach for PMU data recovery." *IEEE Transactions on Smart Grid* 10.4 (2018): 4554-4565.
- [4] Candès, Emmanuel J., and Terence Tao. "The power of convex relaxation: Near-optimal matrix completion." *IEEE transactions on information theory* 56.5 (2010): 2053-2080.
- [5] Ma, Shiqian, Donald Goldfarb, and Lifeng Chen. "Fixed point and Bregman iterative methods for matrix rank minimization." *Mathematical Programming* 128.1 (2011): 321-353.