World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3052

DIMENSIONALITY REDUCTION AND FITTING METHOD FOR HIGH-DIMENSIONAL DATA BASED ON SVD AND LEAST SQUARES—A CASE STUDY OF MINE DATA PROCESSING

JiaYuan Zhang
SWUFE-UD Institute of Data Science at SWUFE, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.
Corresponding Email: 18684018314@163.com

Abstract: In the digital era, the explosive growth of high-dimensional data poses significant challenges to storage, transmission, and computational efficiency. Mine data, characterized by its multi-source heterogeneity, high dynamism, and high dimensionality, presents particularly acute challenges. This paper proposes a novel method for dimensionality reduction and fitting of high-dimensional data by combining Singular Value Decomposition (SVD) with least squares, and demonstrates its first application in mine data processing. The method achieves efficient data compression and precise fitting by extracting principal singular values and vectors via SVD, projecting high-dimensional data into a low-dimensional space, and solving for optimal weight vectors using least squares. A pseudo-inverse is constructed to avoid numerical instability, ultimately completing the fitting of the target dataset. Experimental results show that the method performs exceptionally well in terms of residual distribution, model bias, data noise, and fitting adequacy: residuals approximate a normal distribution, confirming that errors primarily stem from data noise. This study provides a reliable technical pathway for processing high-dimensional mine data, with future optimizations possible through the introduction of noise reduction modules.

Keywords: SVD method; Least squares fitting; Data dimensionality reduction; Mine data processing; Error analysis

1 INTRODUCTION

In the current digital age, data across various fields is growing exponentially, with increasing complexity in dimensionality. While high-dimensional data contains rich information, it also presents formidable challenges, making research into high-dimensional data compression urgent. From a storage perspective, high-dimensional data occupies substantial space, forcing enterprises and institutions to expand hardware infrastructure, thereby driving up costs. In transmission, high-dimensional data requires prolonged transfer times, hindering real-time sharing and interaction. Moreover, high-dimensional data increases computational complexity, reducing the efficiency of data analysis and processing. Traditional algorithms often struggle to handle high-dimensional data, falling short of practical requirements.

The rapid advancement of mine monitoring technologies has also led to the generation of vast amounts of high-dimensional data. While its high resolution, dynamism, and dimensionality support critical tasks like geological modeling and resource assessment, the storage, transmission, and real-time processing of multi-source heterogeneous data remain problematic. The complexity of data fusion escalates computational resource demands, rendering traditional methods inadequate in balancing efficiency and precision.

Vats Deepak compared various common methods in demensionality reduction, mentioned the pros and cons of SVD method[1]. Hastie Trevor combined alternating least squares and SVD method to further solve matrix-completion problem[2]. M.E.Hochstenbach provide a novel method to improve the SVD decomposition efficiency of large matrix[3]. Alkiviadis G. Akritas explained how SVD can be applied to solve least squares problems and data compression[4]. Zhang Chongchong combined NAEEMD and frequency constrained SVD to denoising the mine microseismic signals[5]. Li Shanshan applied SVD to multi-label learning dimensionality reduction, improved classification efficiency[6]. Yang Xinyu innovatively combined K-SVD and SVD for wireless sensor network data, maintaining accuracy while drastically cutting energy consumption[7]. Li Ke enhanced high-dimensional data processing efficiency via an improved randomized SVD algorithm[8]. Zhu Quanjie and Tang Fei employed EMD-SVD and multi-layer SVD, respectively, for denoising mine microseismic signals, significantly improving signal quality[9][10]. These studies demonstrate the efficacy of SVD method in data processing domains. However, in mine data processing, existing research only limits it to denoising. Thus, this study innovatively integrates SVD with least squares, specifically targeting mine data characteristics to address dimensionality reduction and fitting challenges.

This paper’s contributions are: (1)Combining SVD and least squares for high-dimensional data dimensionality reduction and fitting; (2)Pioneering the method’s application in mine data dimensionality reduction; (3)Diagnosing error sources and assessing their impact on fitting results.

2 RELATED THEORIES

Singular Value Decomposition (SVD) is a fundamental data processing technique that extracts principal singular values and their corresponding vectors, projecting high-dimensional data into a low-dimensional space while retaining directions of maximum variance, thereby achieving dimensionality reduction.

For any real or complex matrix$A \in C^{m\times n}$（assuming m≥n）can be broken down into:

$$A = U\sum V^T \tag{1}$$

Where $V \in R^{m\times n}$contains right singular vectors (orthogonal basis of the original feature space).

The diagonal elements of $\sum$,$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ indicate the importance of each direction.

Least squares is a mathematical optimization method for linear regression, minimizing the sum of squared errors to find the best-fitting curve or hyperplane. In matrix form:

$$y = X\beta + \varepsilon \tag{2}$$

Where $y \in R^n$,$X \in R^{n\times(p+1)}$(first column all 1s for intercept β_0),$\beta \in R^{p+1}$ (regression coefficients), and $\varepsilon \in R^n$(error vector).

The optimization problem can be formulated as follows:

$$\min_{\beta} \|y - X\beta\|_2^2 \tag{3}$$

Setting the derivative to zero yields:

$$\frac{\partial}{\partial\beta}\|y - X\beta\|_2^2 =- 2X^T(y - X\beta) = 0 \tag{4}$$

Where X^TX is not singular：

$$\beta = (X^TX)^{-1}X^Ty \tag{5}$$

The geometric significance of least squares is to find the projection of $X\beta$ on the column space of X so that the residuals are orthogonal to the column space:

$$X^T(y - X\beta) =0 \tag{6}$$

When X^TX is invertible ,Pseudo-inversion is required, and the SVD method is used in this paper to avoid direct inversion.

3 EXPERIMENT

Based on the SVD method and the least squares method, the weight vector is found to realize the dimensionality reduction of the original dataset and fit the target dataset as much as possible, and then the residuals are calculated, the residuals are analyzed, the source of the error is confirmed and the advantages and disadvantages of the model are evaluated, and the experimental flow chart is Fig.1.

This flowchart outlines an SVD-based regression modeling workflow: standardizing data, performing SVD decomposition for dimensionality reduction and regression fitting, then calculating predictions and residuals. Error analysis systematically evaluates four aspects: (1) residual distribution (normality/range), (2) model bias (residual-prediction correlation), (3) noise (autocorrelation) and (4) fit adequacy (R²). The integrated process ensures stable high-dimensional computation and reliable modeling through comprehensive diagnostics, providing a complete data-to-evaluation pipeline.

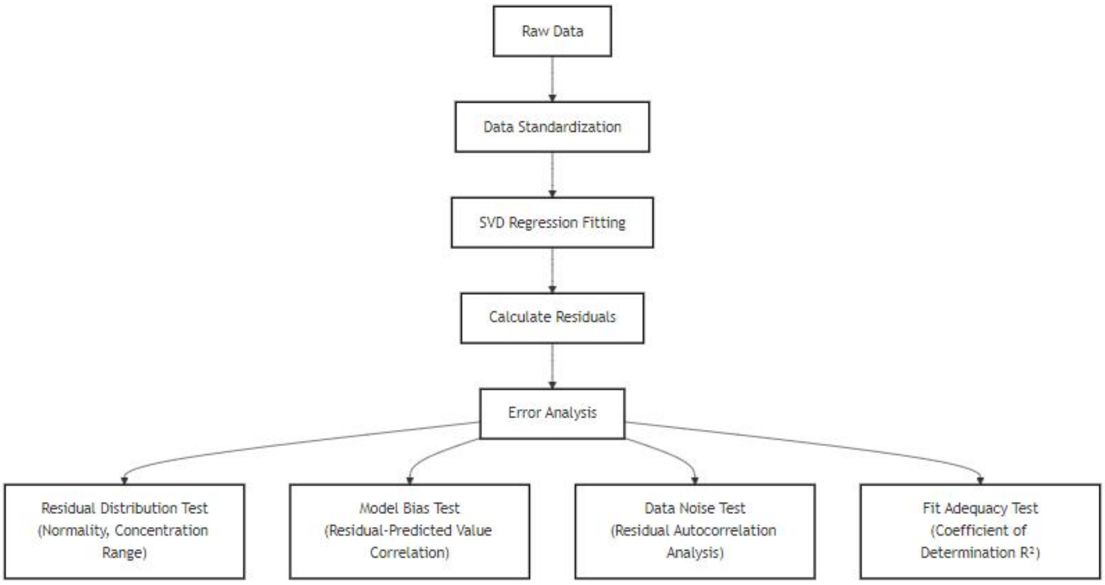
Raw Data
Data Standardization
SVD Regression Fitting
Calculate Residuals
Error Analysis
Residual Distribution Test (Normality, Concentration Range)
Model Bias Test (Residual-Predicted Value Correlation)
Data Noise Test (Residual Autocorrelation Analysis)
Fit Adequacy Test (Coefficient of Determination R²)

Figure 1 Flow Chart of the Experiment

It is known that in order to establish a mathematical model to reduce the dimensionality of a 10000*100 dataset X to a 10000*1 dataset N and fit another 10000*1 dataset y, it is necessary to find the weight vector so that N=Xβ and the residuals are minimized. The solution is to normalize X and y, first decompose SVD to construct a pseudo-inverse, find the least squares solution, and then calculate the residuals. Perform a series of model evaluations and error analyses using residuals and raw data. The specific modeling process is as follows:

Firstly, X and y are standardized, and the data are converted into a distribution with a mean of 0 and a standard deviation of 1 according to columns (features) to eliminate dimensional differences, which is conducive to improving numerical stability.

$$\mathrm{X_{scaled}} = \frac{\mathrm{X}- \mu_{\mathrm{X}}}{\sigma_{\mathrm{X}}} \tag{7}$$

$$\mathrm{y_{scaled}} = \frac{\mathrm{y}-\mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}} \tag{8}$$

Where μ_X, μ_y is the average value of each column,σ_X， σ_y is the standard deviation of each column.

Secondly, the centralized matrix is decomposed.

$$X = U \sum V^T \tag{9}$$

$U \in R^{m\times m}$is the left singular vector matrix, and the column vector is orthogonal.

$\sum \in R^{m\times n}$is a semi-positive definite diagonal matrix, Diagonal elements are singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$, Indicates the importance of each ingredient.

$V \in R^{n\times n}$ is the right singular vector matrix, and the column vector is orthogonal.

Then we need to construct the Moore-Penrose pseudo-inverse X^+ that defines X in the SVD:

$$X^+ = V \sum^+ U^T \tag{10}$$

Where $\sum^+$ is a diagonal matrix, which is obtained by taking the reciprocal of each non-zero element of $\sum$ and transposing it:

$$\sum^+ = \begin{bmatrix} {}^1/_{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {}^1/_{\sigma_n} \end{bmatrix} \tag{11}$$

With pseudo-inverse, the explicit expression of the least squares solution can be expressed as:

$$\beta = X^+ y = V \sum^+ U^T y \tag{12}$$

This method can avoid the numerical instability issues caused by direct inversion. The residuals and MSE are then calculated to assess the model:

$$\varepsilon = y - X\beta \tag{13}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \varepsilon^2 \tag{14}$$

The residuals are computed as the difference between predicted and actual values, and the Mean Squared Error (MSE) is then derived.

Based on the principle of the relevant theory, a series of indicators are obtained from the analysis of four aspects: residual distribution, model bias, data noise and fitting adequacy.

For the residual distribution, this figure should perform as the normal distribution trend, which indirectly justifies the use of least squares.

Then,calculating the Pearson correlation coefficient for the residuals and predicted values to judge the model bias:

$$corr(\varepsilon, y) = \frac{Cov(\varepsilon,y)}{\sigma_\varepsilon \sigma_y} \tag{15}$$

The residual auto-correlation coefficient (calculate the correlation of the residual sequence with its lag k period) can be used for noise diagnostics:

$$ACF(k) = \frac{Cov(\varepsilon_i,\varepsilon_{i-k})}{\sigma_i^2} \tag{16}$$

For underfitting diagnoses, the coefficient of determination is calculated to measure the model's ability to interpret variation in the data:

$$R^2 = 1 - \frac{SSR}{SST} \tag{17}$$

Where SSR (Sum of Squares of Residuals) represents the variation that is not explained by the model, and SST (Sum of Squares of Total Dispersion) represents the total variation of the data.

After the above experiments or derivation or research analysis, a series of relevant conclusions of model diagnosis and error analysis are obtained, which is shown as the following part:

4 RESULTS

As shown in Fig. 2, the residual distribution closely approximates a normal distribution, and the shape of the residual interval distribution line chart is almost completely in line with the theoretical assumptions before modeling, which verifies the rationality of the model. The single-peak, approximately symmetrical distribution indicates that the residuals are mainly dominated by random noise without significant systematic bias, which is consistent with the previous conclusion of "no model bias" based on the correlation between residuals and predicted values. The residuals are centrally distributed in the (-3,3) interval and the tail decays rapidly, which is in line with the expectation of normal error distribution, and is mutually corroborated by the diagnosis of "data noise dominance" in the copy. The stability of the overall distribution morphology further supports the excellent fitting performance reflected by the correlation coefficient, indicating that the feature information retained after the dimensionality reduction of SVD has fully captured the data rules. This distribution characteristic essentially reflects the statistical characteristics of Gaussian noise, and does not require segmentation processing or model correction, which fully satisfies the error assumption of linear regression models.

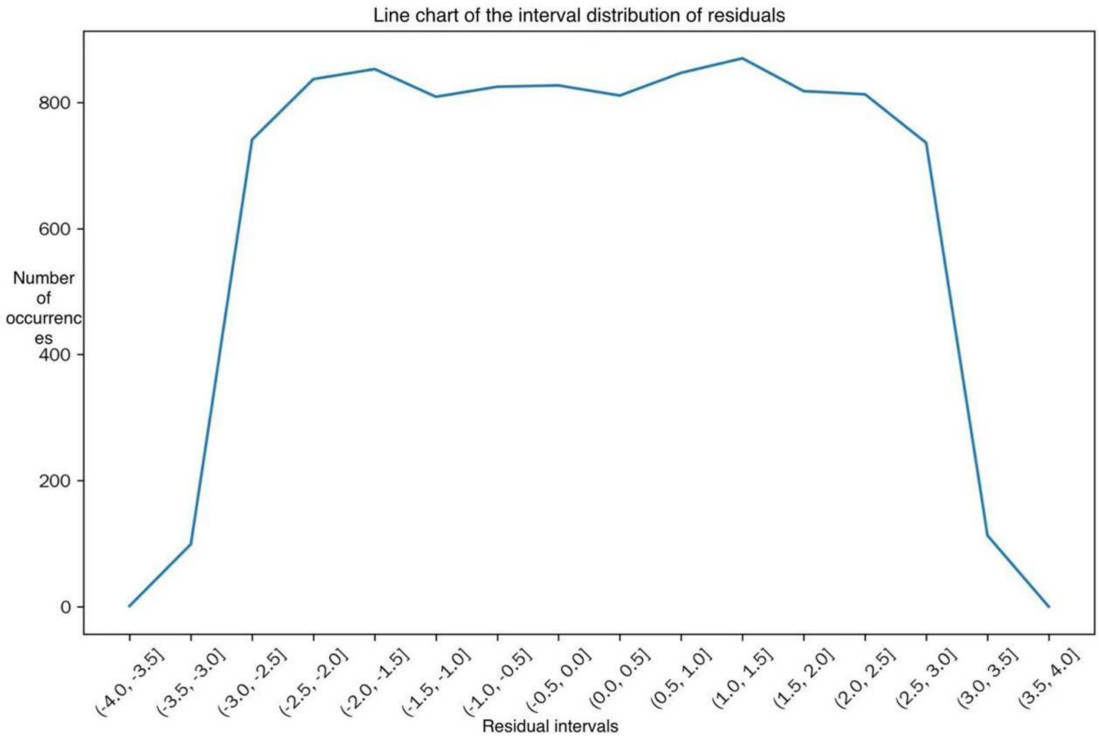

Figure 2 Line Chart of the Interval Distribution of Residuals

The calculated MSE is 3.04, which is relatively small for the original value of 400-500, indicating that the deviation between the predicted value and the actual value is low, and the prediction accuracy of the model is high.

The following is the error analysis data: Model bias: No significant model bias (residuals-predicted value correlation = -0.000), Data Noise: Data Noise Dominates (Residual Auto-correlation=0.004), Underfit: No significant underfit (R^2=0.989)

Here the correlation of the residuals with the predicted values is almost zero, indicating that the model performs well in capturing the linear and nonlinear relationships in the data without significant systematic bias. The current residuals have extremely low auto-correlation, which means that the residuals are more like random noise. It may be that there are many random fluctuations or measurement errors in the data itself, which interfere with the model's learning of real data patterns. This is 0.989, indicating that the model can account for most of the variation in the data without obvious underfitting. This indicates that the features and functions that the model contains are good at capturing patterns in the data.

In summary, the SVD and least squares-based model achieves excellent fitting performance. Error analysis confirms that the primary source of error is data noise.

5 CONCLUSIONS

In this paper, we use the method of combining SVD and least squares method, solve the problem of dimensionality reduction fitting, and verify the performance of the model by model verification and error analysis, and finally establish a model to find the optimal weight vector, and prove the reliability of the model in the field of linear mine data processing. However, this model still have some disadvantages: First, the robustness of the model need to be examined, this method should be suitable and effective for various situation of data. Second, the result of the model fail to match the practical environment that require very high accuracy. In the future, it is expected that a noise reduction module will be added to further reduce the fitting error and improve the model performance. Besides, based on this method, we can add a regularization term to the standard least squares objective function to address instability in overfitting, sick matrices, or high-dimensional data, which will improve the adaptability of this model.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Vats D, Sharma A. Dimensionality Reduction Techniques: Comparative Analysis. Journal of Computational and Theoretical Nanoscience, 2020, 17(6): 2684-2688.

[2] Hastie T, Mazumder R, Lee J D, et al. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. Journal of machine learning research, 2015, 16: 3367-3402.

[3] M E Hochstenbach. Harmonic and Refined Extraction Methods for the Singular Value Problem, with Applications in Least Squares Problems. BIT numerical mathematics, 2004, 44(4): 721-754.

[4] Alkiviadis G A, Gennadi I M. Applications of singular-value decomposition (SVD). Mathematics and Computers in Simulation, 2004, 67(1): 15-31.

[5] Zhang Chongchong, Shi Yannan, Liu Jiangong, et al. A denoising method of mine microseismic signal based on NAEEMD and frequency-constrained SVD. The Journal of Supercomputing, 2022, 78(15): 17095-17113.

[6] Li Shanshan, Tian Wenquan, Pan Zhenggao. Multi-label Learning Algorithm Based on SVD and Kernel Extreme Learning Machine. Journal of Suzhou University, 2020, 35(10): 70-74.

[7] Yang Xinyu, Li Aiping, Duan Liguo, et al. WSN data compression based on dictionary learning and compressed sensing. Computer Engineering and Design, 2022, 43(09): 2448-2455.

[8] Li Ke. Randomized Low-Rank Approximate Algorithms on High Dimensionality Reduction with Applications. China University of Mining and Technology, 2023.

[9] Zhu Quanjie, Sui Longkun, Chen Xuexi, et al. Denoising method and application of mine microseismic signal based on EMD-SVD. Safety and Environmental Engineering, 2024, 31(03): 110-119.

[10] Tang Fei, Liu Zhiwen. Study on multi-layer joint noise of mine microseismic signal. Nonferrous Metals (Mining Section), 2024, 76(04): 92-101.