E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.31 No.3 July - September 2026 Page: [105-115]
Original research article

A Reconstruction Approach for Noisy Data by using Techniques Based on Hankel Matrix and Least-Squares Approximation

Saifon Chaturantabut*

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

Received 4 August 2026; Received in revised form 12 September 2026
Accepted 17 September 2026; Available online 30 September 2026

ABSTRACT

Estimating missing data in volatile time series with oscillatory and noisy patterns is a challenging issue in many applications. This work introduces a technique for missing data imputation that combines data denoising and least-squares approximation. Our method first applies the Hankel Singular Value Decomposition (HSVD) to the dataset to filter out noise and capture the underlying data trend. The cleaned, trend-focused data then serves as the basis for a least-squares approximation to accurately estimate the missing values. This approach is specifically applicable for noisy datasets where related information is available, such as financial data from a stock market. The efficiency of this combined approach is validated through a numerical case study on the stock prices of companies within Thailand's healthcare sector. The results confirm that preprocessing the data with HSVD before applying the least-squares estimation leads to a more accurate and reliable reconstruction of missing entries, highlighting the method's potential for financial time series analysis.

Keywords: Denoising method; Singular values decomposition; Hankel matrix; Financial data

1. Introduction

Missing data is a challenging problem across numerous scientific and engineering disciplines. Missing data arises from a variety of factors including sensor malfunctions, transmission failures, human error, and disparate data collection rates [1]. This inherent imperfection in data acquisition underscores the critical importance of data imputation in the analytical pipeline. Extensive research has been dedicated to the imputation of time series data, often fo-

***Corresponding author:** saifon@mathstat.sci.tu.ac.th doi.org/10.70730/sta.v31i3.260677

cusing on the efficiency of various imputation methods within specific datasets. For instance, studies have explored the application of imputation techniques to traffic data to enhance the predictive capabilities of Intelligent Transportation Systems, which are frequently hampered by data incompleteness [2]. Similarly, investigations utilizing meteorological data have evaluated the performance of both conventional and contemporary imputation methods in time series contexts [3]. Research on air quality data has also addressed imputation, aiming to mitigate bias and improve the efficiency of environmental modeling [4].

The problem of missing data could become difficult to handle in domains where data is often characterized by high volatility, noise, and oscillatory patterns, such as finance data. Stock prices, trading volumes, and other financial indicators are subject to gaps due to various reasons, including market closures, trading halts, data recording errors, or low-frequency trading of certain assets [5, 6]. The proper handling of this missing information is crucial, as many standard statistical and econometric models require complete datasets. The presence of incomplete data can lead to biased analyses, flawed model training, and unreliable forecasts, which makes the development of robust imputation techniques a critical area of research [6, 7].

A wide array of methods has been developed to address missing data, ranging from simple statistical techniques to complex machine learning models. Traditional methods include mean, median, or Last Observation Carried Forward (LOCF) imputation [**?**, **?**, 8, 9]. While these methods are straightforward to implement, they often fail to capture the underlying dynamics of time series data and can introduce significant bias, especially when data is not missing completely at random.More advanced techniques have shown greater promise. Multiple Imputation by Chained Equations (MICE) has been a popular approach, but can be limited in modeling the non-linear relationships often present in financial data [12]. Methods based on the Expectation-Maximization (EM) algorithm have also been widely used. More recently, the field has seen a surge in machine learning and deep learning approaches, including methods based on k-Nearest Neighbors (k-NN), Random Forests, and various neural network architectures like Denoising Autoencoders (DAE) and Generative Adversarial Networks (GANs) [13–15]. While these models are powerful, they often require large amounts of complete data for effective training, which can be a limitation.

Financial time series are notoriously noisy, which can corrupt the data structure and complicate the imputation task. Consequently, a growing body of research suggests that separating the signal from the noise is a crucial preliminary step for effective data analysis and reconstruction. Denoising methods based on signal processing techniques, such as the Wavelet Transform and Empirical Mode Decomposition (EMD), have been successfully combined with forecasting models to improve performance [16].Singular Value Decomposition (SVD) has emerged as a powerful tool for both denoising and data reconstruction. The foundational work by Broomhead and King demonstrated that by constructing a Hankel matrix (a matrix of time-delayed embedding vectors) from a one-dimensional time series, one can use SVD to decompose the signal into a set of orthogonal modes [17]. The dominant singular values typically correspond to the underlying signal trend, while the smaller values represent noise. By truncating these smaller

singular values and reconstructing the signal, a denoised version of the time series can be obtained. This method, often referred to as Hankel-SVD (HSVD), has proven effective in various applications, such as micro-vibration signals [18], partial discharge phenomena [19], and ground-penetrating radar [**?**].

Several studies have explored SVD-based methods for direct imputation, often by framing the problem as a low-rank matrix completion task [21]. However, the performance of these methods can be degraded by high levels of noise. This motivates a two-stage approach, as proposed in this paper, where HSVD is first used explicitly for denoising. This strategy is based on the rationale that a cleaner, trend-focused signal provides a more reliable foundation for subsequent approximation methods. By filtering the noise before imputation, the subsequent least-squares approximation can operate on a more stable representation of the data, allowing it to more accurately model the relationships between observed data points and estimate the missing values. This combined approach leverages the strengths of both signal processing and statistical approximation to tackle the dual challenges of noise and missingness inherent in volatile time series.

2. Denoising Method using Hankel Singular Value Decomposition

This section discusses the construction of a Hankel matrix from noisy time series data and explains how singular value decomposition (SVD) of this matrix can be used to reduce noise.

Let $x(i)$ be the value of the data at time $t_i \in [0,T]$ for $i = 1,...,N$. We can define a vector of one-dimensional discretized data x with N components as $\mathrm{x} = [x(1), x(2), \ldots, x(N)]^T \in \mathbb{R}^N$. In general, the vector x constructed from the original data may contain unwanted noise. The Hankel matrix H corresponding to the time series data x can be generated from the signal x as shown below

$$\mathrm{H} = \begin{bmatrix} x(1) & x(2) & \cdots & x(n) \\ x(2) & x(3) & \cdots & x(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-n+1) & x(N-n+2) & \cdots & x(N) \end{bmatrix},$$

where N is the length of x and the number of column $n \in (1, N)$ is a positive integer that makes the Hankel matrix close to a square matrix as much as possible. That is, if n and m are the numbers of columns and rows for H, respectively, they can be determined as shown below.

$$n = \begin{cases} \frac{N}{2} & , N \text{ is even} \\ \frac{N+1}{2} & , N \text{ is odd} \end{cases}, \qquad (2.1)$$

where $m = N + 1 - n$. Suppose that x consists of the denoised component, denoted by x_s, and the noise component, denoted by w_n. In general, w_n is assumed to be the additive Gaussian noise. Then x can be expressed as:

$$\mathrm{x} = \mathrm{x}_s + \mathrm{w}_n.$$

As a result, if H, H_s and H_{w_n} are the corresponding Hankel matrices of x, x_s and w_n, respectively, then

$$\mathrm{H} = \mathrm{H}_s + \mathrm{H}_{w_n}. \qquad (2.2)$$

SVD will be used to separate the main data trend and the noise subspaces as shown next.

Consider the singular value decomposition of the Hankel matrix $\mathrm{H} \in \mathbb{R}^{m\times n}$:

$$\mathrm{H} = \mathrm{USV}^T,$$

where $\mathrm{U} \in \mathbb{R}^{m\times m}$ and $\mathrm{V} \in \mathbb{R}^{n\times n}$ are unitary matrices, and $\mathrm{S} \in \mathbb{R}^{m\times n}$ is a diagonal matrix that can be expressed as: $\mathrm{S} = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$, with diagonal matrix

$$\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_{R_H}) \in \mathbb{R}^{R_H \times R_H},$$

for $\sigma_1 > \sigma_2 > \ldots > \sigma_{R_H} > 0$, and R_H is the rank of H. Note that U and V are called the left and right singular matrices.

Let k be the dimension of the subspace for columns of H_s corresponding to the original unpolluted signal x_s. Then, the dimension k is less than $R_H = rank(\mathrm{H})$. Therefore, if we partition the matrices U, V, and S in the forms, respectively, as

$$\begin{aligned} \mathrm{U} &= [\mathrm{U}_k, \mathrm{U}_{w_n}], \mathrm{V} = [\mathrm{V}_k, \mathrm{V}_{w_n}], \\ \mathrm{S} &= \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_{w_n} \end{bmatrix}, \end{aligned}$$

where the matrix $\mathrm{U}_k \in \mathrm{R}^{m\times k}$ is the first k columns of U, the matrix $\mathrm{U}_{w_n} \in \mathrm{R}^{m\times(m-k)}$ is the last $m-k$ columns of U, the matrix $\mathrm{V}_k \in \mathrm{R}^{n\times k}$ is the first k columns of V, the matrix $\mathrm{V}_{w_n} \in \mathrm{R}^{n\times(n-k)}$ is the last $n-k$ columns of V, the matrix $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k\times k}$, and the matrix $\Sigma_{w_n} \in \mathbb{R}^{(m-k)\times(n-k)}$ contains the remaining lower-right corner entries in S. We can express the SVD of H as

$$\begin{aligned} \mathrm{H} &= \mathrm{USV}^T \\ &= [\mathrm{U}_k, \mathrm{U}_{w_n}] \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_{w_n} \end{bmatrix} \begin{bmatrix} \mathrm{V}_k^T \\ \mathrm{V}_{w_n}^T \end{bmatrix} \\ &= \mathrm{U}_k\mathrm{S}_k\mathrm{V}_k^T + \mathrm{U}_{w_n}\mathrm{S}_{w_n}\mathrm{V}_{w_n}^T. \end{aligned}$$

Note that, from the above equation, the first part $\mathrm{H}_k := \mathrm{U}_k\mathrm{S}_k\mathrm{V}_k^T$ is expected to be corresponding to the original signal and the second term $\mathrm{U}_{w_n}\mathrm{S}_{w_n}\mathrm{V}_{w_n}^T$ is expected to be corresponding to the noise component. The accuracy of this separation entirely depends on the choice of the dimension k. When the value of k is too small, the resulting denoised data may ignore the important feature of the actual signal. However, when the value of k is too large, the resulting denoised signal may include also some part of the noise component. Ideally, we would like to choose k so that H_k is the same as H_s from (2.2) to recover the actual trend of the time series data. The appropriate choice of k can be observed from the decay of singular values in the matrix S.

The denoised data is finally recovered from the matrix H_k with the same position as define for the Hankel matrix H in (2.1). That is, if x_k is defined to be the approximated denoised data, then

$$\mathrm{x}_k = [x_k(1), x_k(2), \ldots, x_k(N)]^T \quad (2.3)$$

with the components $x_k(1), x_k(2), \ldots, x_k(N)$ extracted from the first row and last column of H_k as shown below:

$$\mathrm{H}_k = \begin{bmatrix} x_k(1) & x_k(2) & \cdots & x_k(n) \\ x_k(2) & x_k(3) & \cdots & x_k(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ x_k(N-n+1) & x_k(N-n+2) & \cdots & x_k(N) \end{bmatrix}.$$

The steps for denoising time series data described above are summarized in Algorithm 1.

Algorithm 1 Algorithm for denoising time series data using HSVD

Input: $\mathrm{x} = [x(1), x(2), \ldots, x(N)]^T$

1: Form a Hankel matrix H for x
2: Find SVD of H: $\mathrm{H} = \mathrm{USV}^T$
3: Select dimension k.
4: Form
5: $\quad \mathrm{U}_k := \mathrm{U}(:, 1:k)$,
6: $\quad \mathrm{V}_k := \mathrm{V}(:, 1:k)$ and
7: $\quad \mathrm{S}_k := \mathrm{S}(1:k, 1:k)$.
8: Construct $\mathrm{H}_k = \mathrm{U}_k\mathrm{S}_k\mathrm{V}_k^T$.
9: Reconstruct denoised data:
10: $\quad \mathrm{x}_k$ in 2.3 from H_k.

Output : Denoised Data:
$\mathrm{x}_k = [x_k(1), x_k(2), \ldots, x_k(N)]^T$

3. Missing Data reconstruction

This section describes how least-squares approximation is used to reconstruct missing components of incomplete data. We will approximate an incomplete

data vector by projecting on a subspace spanned by a basis that represents other related complete data vectors.

Let $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}\}$ be a complete data set. Define $\mathcal{Y} := \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}\} \subset \mathbb{R}^n$ and $\mathrm{Y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}] \in \mathbb{R}^{N \times n_s}$. Suppose $\widehat{\mathrm{y}}$ is an incomplete sample that has n_c known components and $n_g = N - n_c$ unknown components. Let $C := \{c_1, c_2, \ldots, c_{n_c}\} \subset \{1, 2, \ldots, N\}$ be the indices of the known components in $\widehat{\mathrm{y}}$ and define $\mathrm{C} = [\mathrm{e}_{c_1}, \ldots, \mathrm{e}_{c_{n_c}}] \in \mathbb{R}^{N \times n_c}$, where $\mathrm{e}_{c_i} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ is the c_i-th column of the identity matrix $\mathrm{I}_N \in \mathbb{R}^{N \times N}$, for $i = 1, \ldots, n_c$. Note that, pre-multiplying C^T is equivalent to extracting the n_c rows corresponding to the indices $c_1, \ldots, c_{n_c}$. Similarly, let $\mathcal{G} = \{g_1, g_2, \ldots, g_{n_g}\} \subset \{1, 2, \ldots, N\}$ be the indices of the unknown components in $\widehat{\mathrm{y}}$ and define $\mathrm{G} = [\mathrm{e}_{g_1}, \ldots, \mathrm{e}_{g_{n_g}}] \in \mathbb{R}^{N \times n_g}$. Let $\widehat{\mathrm{y}}_c := \mathrm{C}^T\widehat{\mathrm{y}} \in \mathbb{R}^{n_c}$ and $\widehat{\mathrm{y}}_g := \mathrm{G}^T\widehat{\mathrm{y}} \in \mathbb{R}^{n_g}$. Then, the known components and the unknown components are contained in the vectors $\widehat{\mathrm{y}}_c$ and $\widehat{\mathrm{y}}_g$, respectively. The missing components contained in $\widehat{\mathrm{y}}_g$ will be approximated by first projecting $\widehat{\mathrm{y}}$ onto the column span of a basis matrix U with rank r, where $r \le R, R := \mathrm{rank}(\mathrm{Y})$. I.e.

$$\widehat{\mathrm{y}} \approx \mathrm{U}\mathrm{a},$$

or

$$\widehat{\mathrm{y}}_c \approx \mathrm{U}_c\mathrm{a} \text{ and } \widehat{\mathrm{y}}_g \approx \mathrm{U}_g\mathrm{a},$$

for some coefficient vector $\mathrm{a} \in \mathbb{R}^r$, and where $\mathrm{U}_c := \mathrm{C}^T\mathrm{U} \in \mathbb{R}^{n_c \times r}$, $\mathrm{U}_g := \mathrm{G}^T\mathrm{U} \in \mathbb{R}^{n_g \times r}$. Note that the choice of the basis matrix U will be discussed in Section 3.1.

The known components contained in $\widehat{\mathrm{y}}_c$ are then used to determine the coefficient vector a through the approximation $\widehat{\mathrm{y}}_c \approx \mathrm{U}_c\mathrm{a}$ from the following least-squares problem:

$$\min_{\mathrm{a} \in \mathbb{R}^r} \|\widehat{\mathrm{y}}_c - \mathrm{U}_c\mathrm{a}\|_2^2. \tag{3.1}$$

The solution obtained from the corresponding normal equation of the above problem is given by $\mathrm{a} = \mathrm{U}_c^+\widehat{\mathrm{y}}_c$, where $\mathrm{U}_c^+ = (\mathrm{U}_c^T\mathrm{U}_c)^{-1}\mathrm{U}_c^T$ is the Moore–Penrose pseudoinverse. That is,

$$\widehat{\mathrm{y}}_g \approx \mathrm{U}_g\mathrm{a} = \mathrm{U}_g\mathrm{U}_c^+\widehat{\mathrm{y}}_c. \tag{3.2}$$

The steps described above are summarized in Algorithm 2.

Algorithm 2 Standard least-squares approach for approximating missing data

Inputs:

- Complete snapshot set $\{\mathrm{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^N$ and, dimension $r \le rank(\{\mathrm{y}_j\}_{j=1}^{n_s})$
- Incomplete data $\widehat{\mathrm{y}} \in \mathbb{R}^N$ with known entries in $\widehat{\mathrm{y}}_c \in \mathbb{R}^{n_c}$ and unknown entries in $\widehat{\mathrm{y}}_g \in \mathbb{R}^{n_g}$, where $n_c + n_g = N$.

1: Create snapshot matrix: $\mathrm{Y} = [\mathrm{y}_1, \ldots, \mathrm{y}_{n_s}] \in \mathbb{R}^{N \times n_s}$ and let $R = rank(\mathrm{Y})$.
2: Construct basis U of rank $r \le R$ for Y.
3: Find coefficient vector a from $\widehat{\mathrm{y}}_c$ using least-squares problem in (3.1):
4: $\min_{\mathrm{a} \in \mathbb{R}^r} \|\widehat{\mathrm{y}}_c - \mathrm{U}_c\mathrm{a}\|_2^2$.
5: Compute the approximation $\widehat{\mathrm{y}}_g \approx \mathrm{U}_g\mathrm{a}$.

Output: Approximation of $\widehat{\mathrm{y}}_g$

From Algorithm 2, a crucial factor to obtain missing data approximation is to have a basis U in Step 2. This basis should capture the main features of the snapshot set $\{\mathrm{y}_j\}_{j=1}^{n_s}$. A common approach to construct this basis optimally will be discussed next.

3.1 Optimal basis for data set

Consider a set of snapshots $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}\}$ where $\mathrm{y}_j \in \mathbb{R}^N$, $j = 1, 2, \ldots, n_s$. Suppose we want to approximate a snapshot y_j by using a set of orthonormal vectors $\{\phi_1, \phi_2, \ldots, \phi_r\} \subset \mathbb{R}^N$, where $r < N$. Then the approximation is in the form $\mathrm{y}_j \approx \sum_{i=1}^r w_i\phi_i$, for some constants w_i, $i = 1, 2, ..., r$.

Alternatively, we can write this approximation in matrix form as $\mathrm{y}_j \approx \Phi_r \mathrm{w}$, where $\Phi_r = [\phi_1, \phi_2, \ldots, \phi_r] \in \mathbb{R}^{N\times r}$ is a matrix of basis vectors and $\mathrm{w} = [w_1, w_2, \ldots, w_r]^T \in \mathbb{R}^r$ is the vector of unknown coefficients. To find w, we use the fact that Φ_r has orthonormal columns, i.e. $\Phi_r^T\Phi_r = \mathrm{I}_r$, and the minimum error occurs when the residual is orthogonal to the column span of Φ_r, i.e. $\Phi_r^T(\mathrm{y}_j - \Phi_r \mathrm{w}) = 0$, which implies that $\mathrm{w} = \Phi_r^T \mathrm{y}_j$ and the approximation becomes

$$\mathrm{y}_j \approx \Phi_r \Phi_r^T \mathrm{y}_j. \tag{3.3}$$

The goal is to find the orthonormal basis that minimizes this approximation error in 2-norm for a given basis rank $r \le R := \mathrm{rank}(\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}\})$, i.e.

$$\min_{\Phi_r \in \mathbb{R}^{N\times r}} \sum_{j=1}^{n_s} \|\mathrm{y}_j - \Phi_r \Phi_r^T \mathrm{y}_j\|_2^2 \tag{3.4}$$

such that $\Phi_r^T\Phi_r = \mathrm{I}_r$, where $\mathrm{I}_r \in \mathbb{R}^{r\times r}$ is the identity matrix. This optimal basis is often called proper orthogonal decomposition (POD), which is also known by other names, for example, Karhunen-Loève decomposition (KLD), principal component analysis (PCA), or singular value decomposition (SVD).

It can be shown [22,23] that POD basis defined above can be obtained from the left singular vector of the snapshot matrix $\mathrm{Y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{n_s}]$. Let $\mathrm{Y} = \mathrm{U}\Sigma\mathrm{V}^T$ be the singular value decomposition of Y, where matrices $\mathrm{U} = [\mathrm{u}_1, \ldots, \mathrm{u}_R] \in \mathbb{R}^{N\times R}$ and $\mathrm{V} = [\mathrm{v}_1, \ldots, \mathrm{v}_R] \in \mathbb{R}^{n_s\times R}$ are matrices with orthonormal columns and $\Sigma = diag(\sigma_1, \ldots, \sigma_R) \in \mathbb{R}^{R\times R}$ is a diagonal matrix with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_R > 0$. Then the POD basis matrix of dimension r is $\mathrm{U}_r = [\mathrm{u}_1, \ldots, \mathrm{u}_r] \in \mathbb{R}^{N\times r}$, $r \le R$, i.e. $\mathrm{U}_r = \arg\min_{\Phi_r \in \mathbb{R}^{N\times r}} \sum_{j=1}^{n_s} \|\mathrm{y}_j - \Phi_r\Phi_r^T\mathrm{y}_j\|_2^2$ with $\mathrm{U}_r^T\mathrm{U}_r = \mathrm{I}_r$. It is well-known [22, 23] that this minimum error is given by

$$\sum_{j=1}^{n_s} \|\mathrm{y}_j - \mathrm{U}_r\mathrm{U}_r^T\mathrm{y}_j\|_2^2 = \sum_{\ell=r+1}^{R} \sigma_\ell^2, \tag{3.5}$$

which is the sum of the neglected singular values $\sigma_{r+1}, \ldots, \sigma_R$ from the SVD of Y.

4. Numerical Results

This section evaluates the effectiveness of the proposed approach for estimating missing data in highly volatile and oscillatory time series data. A data set of stock prices is used as a case study in this work. This case study utilizes daily stock price data from 10 companies in the Thai healthcare sector, collected from April 1, 2023, to April 1, 2024. The companies are listed in Table 1. The inherent volatility of this financial data is evident in Fig. 1, which plots the original stock prices for Bangkok Dusit Medical Services (BDMS). A comparative analysis is performed for the missing data approximation accuracy achieved by applying a least-squares method to the original raw data versus the denoised data.

Ticker	Full Name
BCH	BANGKOK CHAIN HOSPITAL
BDMS	BANGKOK DUSIT MEDICAL SERVICES
BH	BUMRUNGRAD HOSPITAL
CHG	CHULARAT HOSPITAL
EKH	EKACHAI MEDICAL CARE
PR9	PRARAM 9 HOSPITAL
PRINC	PRINCIPAL CAPITAL
SKR	SIKARIN
THG	THONBURI HEALTHCARE GROUP
VIBHA	VIBHAVADI MEDICAL CENTER

Table 1. The companies in healthcare sector considered in the case study.

In Section 4.1, we first employ the HSVD denoising technique described in Section 2 to extract the data trend. Then, Section 4.2 demonstrates the accuracy results of missing data approximation by us-

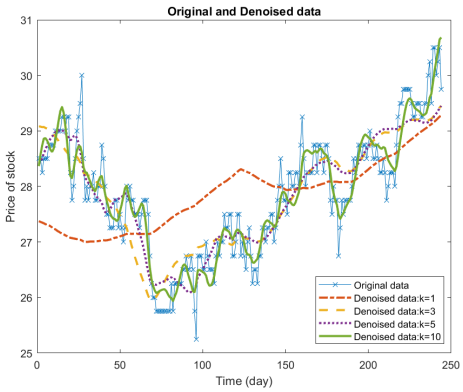


Fig. 1. BDMS stock price and its denoised data using HSVD approach with $k = 1, 3, 5, 10$.

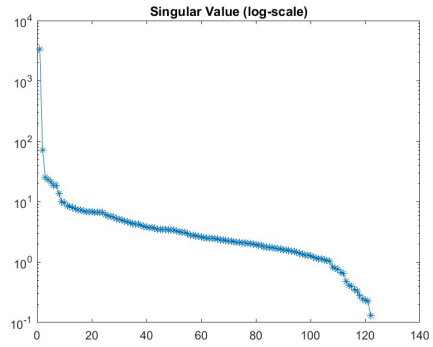


Fig. 2. Singular values of Hankel matrix for the BDMS stock prices.

ing the least-squares approach described in Section 3 based on both the original raw data and the denoised data from Section 4.1.

4.1 Numerical results on HSVD denoising method

To extract the underlying data trend, we employ the HSVD denoising technique described in Section 2. The choice of the dimension k in the HSVD method is crucial; a small k captures only the principal trend, whereas a large k may retain too much of the original noise, as shown in Fig. 1. The selection of an appropriate k can be guided by examining the singular value decay of the corresponding Hankel matrix. As shown for the BDMS stock data in Fig. 2, a sharp drop occurs around the fifth singular value, suggesting that $k = 5$ is a reasonable choice to capture the data's overall trend without overfitting to noise. Fig. 3 illustrates the resulting trends for all 10 companies, which suggest the HSVD approach to take parameter $k = 5$.

4.2 Missing data reconstruction

We now assess the accuracy of the data reconstruction. The analysis compares the performance of the least-squares approximation under two distinct scenarios:

- Case 1: The least-squares method is applied directly with the original, noisy stock price data.
- Case 2: The least-squares method is applied with the denoised data obtained from the HSVD preprocessing step.

Recall that the least-squares approximation given in Section 3 requires to select the low dimension r, which can be done by observing the decay of the singular values as shown in Fig. 4. Fig. 4 suggests that reasonable choices of r are around $r = 2, 3, 4$ where the singular values start to stop decreasing.

The accuracy is quantified using the Root Mean Squared Error (RMSE), Relative RMSE (Rel RMSE) and Mean Absolute Percentage Error (MAPE), as defined in Eqs. (4.1)-(4.3), respectively.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \quad (4.1)$$

$$\text{Rel RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2} \quad (4.2)$$

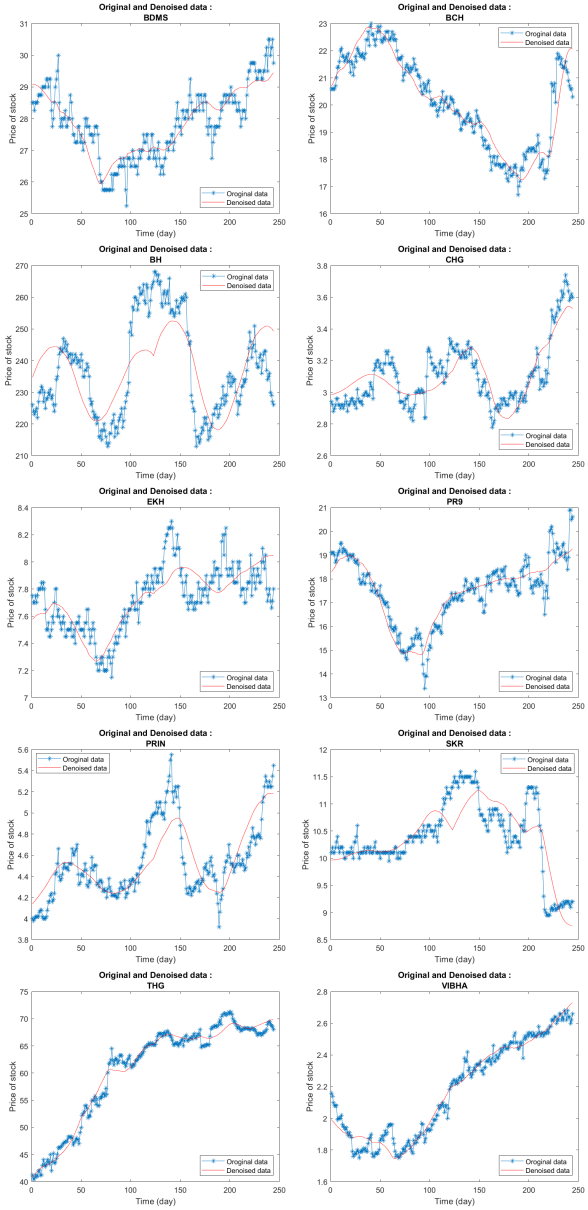


Fig. 3. The actual stock prices (in blue color) of the 10 companies from healthcare sector in Thailand and the corresponding trend (in red color) from the HSVD approach with $k = 5$.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100. \quad (4.3)$$

The experiments were conducted by simulating missing data points and comparing the estimated values to the exact values. Tables 2-4 present the approximation results for three randomly chosen missing components (indices 10, 20, and 30) for both Case 1 and Case 2.

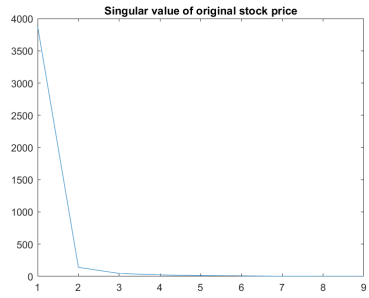


Fig. 4. Singular values of Hankel matrix for the BDMS stock prices.

Table 2. Approximations of three missing components that are randomly chosen and the corresponding RMSE and Rel RMSE when $k = 3$ is used in HSVD and $r = 3$ is used in the least-squares method.

Index	Exact	Case 1: Approx	Case 2: Approx (denoised)
10	2.0000	1.7787	1.9518
20	1.8900	1.7764	1.9444
30	1.7900	1.8124	1.9064
Error: RMSE	-	0.1442	0.0792
Error: Rel RMSE	-	0.0730	0.0434
Error: MAPE	-	6.109%	3.930%

Table 3. Approximations of three missing components that are randomly chosen and the corresponding RMSE and Rel RMSE when $k = 3$ is used in HSVD and $r = 4$ is used in the least-squares method.

Index	Exact	Case 1: Approx	Case 2: Approx (denoised)
10	2.0000	1.9150	1.9531
20	1.8900	1.9181	1.8243
30	1.7900	1.8567	1.8021
Error: RMSE	-	0.0508	0.0471
Error: Rel RMSE	-	0.0273	0.0245
Error: MAPE	-	2.571%	2.165%

Table 4. Approximations of three missing components that are randomly chosen and the corresponding RMSE and Rel RMSE when $k = 5$ is used in HSVD and $r = 3$ is used in the least-squares method.

Index	Exact	Case 1: Approx	Case 2: Approx (denoised)
10	2.0000	1.7787	1.9832
20	1.8900	1.7764	1.9135
30	1.7900	1.8124	1.8629
Error: RMSE	-	0.1442	0.0454
Error: RMSE Rel	-	0.0730	0.0251
Error: MAPE	-	6.109%	2.072%

Table 5. RMSE and Rel RMSE for the approximations of missing data blocks (index ranges 10-15 and 10-20) when $k = 5$ is used in HSVD and $r = 4$ is used in the least-squares method.

Index Range	Error Case 1: Approx		Error Case 2: Approx (denoised)	
	RMSE	Rel. RMSE	RMSE	Rel. RMSE
10-15	0.1527	0.0769	0.0147	0.0114
10-20	0.1362	0.0692	4.7731e-04	0.0101

which are called "Approx" and "Approx (denoised)," respectively in these tables. Note that the missing components are in one of the ten companies' stock prices and the remaining 9 companies contain complete datasets. Table 2 shows the results using $k = 3$ for HSVD and $r = 3$ for the least-squares method. Table 3 demonstrates that increasing the least-squares dimension to $r = 4$ (while keeping $k = 3$) improves the accuracy for both cases, as indicated by lower RMSE and Rel RMSE, and MAPE values. Table 4 shows that further increasing the HSVD parameter to $k = 5$ (while keeping $r = 3$ as in Table 2) yields a slight improvement in the accuracy for Case 2 when compared with the results in Table 2. Note that the change in dimension k used in HSVD has no effect on the results in Case 1 because HSVD is not applied in Case 1. A crucial and consistent observation across all experiments is that Case 2 provides more accurate approximations than Case 1. This highlights the significant benefit of using denoised data for the least-squares approximation.

To further investigate the robustness of our approach, we estimated consecutive missing data blocks. Table 5 details the results for missing data in index ranges 10-15 and 10-20. The results confirm our previous findings that applying the least-squares method to denoised data (Case 2) yields substantially more accurate reconstructions compared to using the original noisy data (Case 1).

Notice that utilizing the HSVD denoising technique as a preprocessing step provides a more stable and reliable background for the least-squares approximation, delivering a promising and accurate method for predicting missing data in noisy time series.

5. Conclusion

This work addresses the challenging problem of missing data imputation in volatile, oscillatory, and noisy time series by introducing a combined approach of Hankel Singular Value Decomposition (HSVD) for denoising and least-squares approximation. This approach is particularly applicable for the datasets where related time series information is available. The numerical results from a case study on stock prices of companies in Thailand's healthcare sector consistently demonstrate the superior performance of the proposed method. Specifically, preprocessing the data with HSVD to extract the underlying trend before applying least-squares estimation significantly enhances the accuracy and reliability of reconstructing missing entries. This is evidenced by lower RMSE, Rel RMSE and MAPE values in "Case 2" (least-squares method with denoised data) compared to "Case 1" (least-squares method with raw original data), even when considering randomly chosen single missing components and consecutive missing data

ranges. These results emphasize the significant potential of this combined approach for robust and accurate missing data imputation in financial time series analysis.

Possible future directions on enhancing the proposed missing data imputation technique include exploring adaptive methods for selecting the optimal 'k' dimension in HSVD, extending its application to a wider range of financial instruments and global markets, and evaluating its effectiveness in handling longer and irregularly spaced missing data segments. Additionally, a comprehensive comparison with other state-of-the-art imputation techniques will be conducted, alongside an investigation into how various data characteristics influence performance. Finally, integrating this method into real-time systems for financial data processing and forecasting could further extend the proposed approach in practical applications.

6. Acknowledgements

The author gratefully acknowledge the financial support provided by Faculty of Science and Technology, Thammasat University, Contract No. SciGR 10/2567.

References

[1] Pigott TD. A review of methods for missing data. Educ Res Eval. 2001;7(4):353-83.

[2] Wu H, Xu L. Imputation methods used in missing traffic data: a literature review. In: ISICA 2019: Communications in Computer and Information Science. Vol. 1205. Springer; 2020.

[3] Yozgatligil C, Aslan S, Iyigun C, Batmaz I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol. 2013;112:143-67.

[4] Wijesekara WMLKNLL, Liyanage L. Comparison of imputation methods for missing values in air pollution data: case study on Sydney Air Quality Index. In: Advances in Intelligent Systems and Computing. Vol. 1130. Springer; 2020.

[5] Kofman P, Sharpe IG. Using multiple imputation in the analysis of incomplete observations in finance. J Financ Econom. 2003;1(2):216-49.

[6] Bryzgalova S, Lerner S, Lettau M, Pelger M. Missing financial data. Rev Financ Stud. 2025;38(3):803-82.

[7] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147-77.

[8] Purwar A, Singh SK. Empirical evaluation of algorithms to impute missing values for financial dataset. In: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). IEEE; 2014. p. 652-6.

[9] Fujimoto S, Mizuno T, Ishikawa A. Interpolation of non-random missing values in financial statements' big data using CatBoost. J Comput Soc Sci. 2022;5(2):1281-301.

[10] Khan D, Lazar A. A comparative study of imputation methods for time series data. In: The International FLAIRS Conference Proceedings. 2023;36(1).

[11] Goldani M. Comparative analysis of missing values imputation methods: a case study in financial series (S&P500 and Bitcoin value data sets). Iran J Finance. 2024;8(1):47-70.

[12] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67.

[13] Gondara L, Wang K. MIDA: multiple imputation using denoising autoencoders. In: Advances in Knowledge Discovery

and Data Mining. Cham: Springer International Publishing; 2018. p. 260-72.

[14] Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006-2017). Artif Intell Rev. 2020;53:1487-509.

[15] Savarimuthu N, Karesiddaiah S. An unsupervised neural network approach for imputation of missing values in univariate time series data. Concurr Comput Pract Exp. 2021;33(9):e6156.

[16] Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS One. 2017;12(7):e0180944.

[17] Broomhead DS, King GP. Extracting qualitative dynamics from experimental data. Physica D. 1986;20(2-3):217-36.

[18] He ZJ, et al. Signal denoising method using AIC-SVD and its application to micro-vibration in reaction wheels. Sensors. 2019;19(22):5032.

[19] Govindarajan S, Subbaiah J, Cavallini A, Krithivasan K, Jayakumar J. Development of Hankel-SVD hybrid technique for multiple noise removal from PD signature. IET Sci Meas Technol. 2019;13(8):1075-84.

[20] Xue W, Luo Y, Yang Y, Huang Y. Noise suppression for GPR data based on SVD of window-length-optimized Hankel matrix. Sensors. 2019;19(17):3807.

[21] Blázquez-García A, Wickstrøm K, Yu S, Mikalsen KØ, Boubekki A, Conde A, et al. Selective imputation for multivariate time series datasets with missing values. IEEE Trans Knowl Data Eng. 2023;35(9):9490-501.

[22] Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika. 1936;1(3):211-8.

[23] Volkwein S. Proper orthogonal decomposition: theory and reduced-order modelling. Lecture notes. Konstanz: University of Konstanz; 2013.