AIMS Mathematics, 11(11): 35929–35956.
DOI: 10.3934/math.20261393
Received: 03 August 2026
Revised: 30 September 2026
Accepted: 14 October 2026
Published: 02 November 2026

https://www.aimspress.com/journal/Math

Research article

Multi-indicator stock clustering based on LSPIA-B-spline control-coefficient representation

Siyu Shen, Lijuan Hu* and Chengzhi Liu

School of Mathematics and Data Science, Hunan College of Humanities, Science and Technology, Loudi 417000, China

* **Correspondence:** Email: 3464@huhst.edu.cn, 1023236794@qq.com.

Abstract: To address the high dimensionality of long stock time series and the limited descriptive capability of a single indicator, this paper proposes a multi-indicator stock-behavior clustering method based on a least-squares progressive iterative approximation (LSPIA)-B-spline control-coefficient representation. First, four financial indicators are constructed from the perspectives of periodic returns, risk exposure, and trading activity, and are standardized in a unified manner. Second, a common cubic B-spline basis is fitted to each stock's standardized multi-indicator series by LSPIA, yielding comparable, low-dimensional control-coefficient matrices without directly solving the least-squares normal equations. Rather than treating the coefficients as ordinary compressed features, we define a distance between the fitted series. On this basis, K-means is used to identify groups of stocks with similar market behavior. The clustering results are comprehensively evaluated using internal clustering indices, clustering stability, and economic-characteristic analysis. The proposed method provides an analytical framework for identifying stock-behavior structures that jointly considers temporal patterns, computational efficiency, and interpretability.

Keywords: stock-behavior clustering; LSPIA; B-spline; functional distance; K-means
Mathematics Subject Classification: 62H30, 65D07, 65D10

1. Introduction

The sectoral structure of a stock market reflects not only the industrial attributes and common fundamentals of listed companies, but also similarities in return dynamics, risk exposure, and trading behavior. Although stocks belonging to the same industry may share certain economic characteristics, their market behavior can exhibit substantial heterogeneity over time. Identifying latent groups of stocks with similar dynamic characteristics is therefore important for understanding market structure, analyzing sector rotation, monitoring financial risk, and supporting portfolio diversification.

Stock clustering provides a data-driven approach to identifying such relationships. Mantegna constructed a distance measure based on correlations between daily logarithmic returns and revealed a hierarchical structure of financial markets through hierarchical clustering [1]. Lao and Shao investigated dynamic clustering based on industry-index returns, illustrating how clustering can characterize relationships among industries and their temporal evolution [2]. These studies established an important foundation for analyzing the endogenous structure of financial markets through similarities in observed market behavior. However, the effectiveness of stock clustering depends critically on how financial time series are represented and how similarities between stocks are measured.

From a data-structural perspective, daily stock-market observations constitute multivariate time series characterized by high dimensionality, temporal dependence, and substantial local fluctuations. As the numbers of stocks, observation periods, and financial indicators increase, directly clustering the original observations may lead to considerable computational costs and sensitivity to short-term fluctuations. Existing time-series clustering methods are commonly categorized into raw-data-based, feature-based, and model-based approaches [3, 4]. Raw-data-based methods calculate similarities directly from the observed series, whereas feature-based and model-based methods first construct compact representations and subsequently perform clustering in the resulting feature or parameter space. The latter approaches provide a natural way to reduce the dimensionality of long financial time series while retaining information relevant to their temporal behavior.

Despite substantial progress in financial time-series clustering, several issues remain worthy of further investigation.

First, an appropriate representation of stock-market behavior should account for multiple aspects of financial dynamics. Traditional approaches frequently rely on stock prices, returns, or their correlation structures. However, stocks with similar return patterns may differ considerably in volatility, liquidity, and trading activity. Financial time series also exhibit characteristics such as volatility clustering, heavy tails, and heterogeneous dependence structures. Recent studies have incorporated volatility-related characteristics, conditional moments, and dynamic dependence into financial clustering, illustrating the value of describing financial assets from multiple perspectives [5, 6]. In addition, Kahssay and Miah investigated stock-price forecasting under different volatility regimes and historical calibration windows, highlighting the relevance of volatility heterogeneity in financial modeling [7]. Although their study addresses forecasting rather than clustering, it provides further motivation for jointly considering return dynamics and volatility characteristics. Accordingly, this paper constructs four financial indicators, namely the 20-day cumulative logarithmic return, the 20-day average rate of change in trading volume, the turnover rate, and the 20-day rolling volatility, to characterize stock-market behavior in terms of periodic returns, risk fluctuations, and trading activity.

Second, the high dimensionality of multivariate financial time series calls for a compact representation that preserves their principal temporal patterns. Directly calculating distances between observations over long periods may increase computational complexity and amplify the influence of local fluctuations. Recent research on multivariate time-series clustering has therefore emphasized extracting representative features or dynamic-dependence information before clustering [8,9]. Among functional representation methods, B-splines provide a flexible means of approximating temporal trajectories through a relatively small number of basis coefficients. D'Urso et al. demonstrated the feasibility of clustering time series using B-spline-based representations [10]. Related studies on

China's stock market have also employed autoencoding techniques to extract low-dimensional latent factors from high-dimensional firm characteristics [11]. These studies suggest that dimensionality reduction can facilitate subsequent financial-data analysis. Nevertheless, constructing a compact representation that simultaneously retains the temporal characteristics of multiple financial indicators and ensures comparability across stocks remains an important consideration.

Third, the distance measure used for clustering should be consistent with the underlying representation. Conventional coefficient-based clustering often calculates distances directly between the coefficients of fitted models. For B-spline representations, however, the Euclidean distance between raw coefficient vectors does not generally coincide with the distance between their corresponding fitted trajectories. This discrepancy arises because B-spline basis functions have local support and overlap, so that coefficient differences contribute unequally to the resulting functions. Functional clustering provides an alternative by defining similarities in terms of fitted trajectories rather than raw model parameters. In related research, Luan and Hamp developed a sliced Wasserstein K-means approach for market-regime classification using multivariate financial time series [12]. Their approach focuses on distributional differences between market-regime windows, whereas the present study focuses on similarities between individual stocks represented by fitted financial-indicator trajectories. These approaches therefore differ in their clustering objects and underlying distance geometries. For the present setting, an important methodological issue is how to construct a distance that reflects differences between fitted B-spline trajectories while remaining compatible with efficient Euclidean clustering algorithms.

B-spline approximation provides a suitable foundation for addressing these issues. Owing to their local support, controllable continuity, and approximation properties, B-splines can represent complex temporal patterns using relatively few control coefficients [13]. When all stocks share a common time parameterization, knot vector, spline degree, and number of control coefficients, their fitted representations possess a unified functional meaning and can be compared consistently across stocks.

To obtain these representations, this paper adopts least-squares progressive iterative approximation (LSPIA), a geometric iterative method for B-spline least-squares fitting [14]. Unlike direct approaches that explicitly solve the associated normal equations, LSPIA progressively updates the control coefficients using the fitting residuals. Its mathematical formulation is closely related to gradient-based optimization, providing a theoretical basis for convergence analysis and parameter selection. Lin et al. reviewed the fundamental theory, convergence properties, and applications of progressive iterative approximation [15], while Rios and Jüttler further investigated the relationships among LSPIA, gradient descent, and stochastic gradient descent [16]. Subsequent studies have explored several strategies to improve the computational efficiency of LSPIA, including hyperpower iterations [17], asynchronous updating [18], randomized iterative schemes [19, 20], and momentum-accelerated randomized methods [21]. In the present study, LSPIA is employed as an iterative fitting mechanism rather than as a newly proposed optimization algorithm. Its role is to obtain compact B-spline representations of multivariate stock-market behavior without explicitly computing the inverse of the associated normal-equation matrix.

Building upon this representation, the relationship between coefficient-space distances and fitted-series distances can be established through the B-spline collocation matrix. Specifically, the squared Frobenius distance between two fitted multivariate trajectories can be expressed as a quadratic form involving the differences between their control-coefficient matrices and the Gram matrix of the

common collocation matrix. By introducing an appropriate matrix factorization, this functional distance can be equivalently represented as the squared Euclidean distance between transformed coefficient features. Consequently, the clustering procedure can operate in a Euclidean feature space while preserving the distance structure induced by the fitted B-spline trajectories. This construction provides a direct connection between functional representation and K-means clustering, rather than treating dimensionality reduction and distance calculation as two independent procedures.

Based on the above considerations, this paper develops a multi-indicator stock-behavior clustering framework integrating B-spline least-squares fitting, functional-distance construction, and K-means clustering. First, four financial indicators are constructed and standardized to characterize the return, risk, and trading characteristics of each stock. Second, a common cubic B-spline basis is employed to fit the multivariate time series through LSPIA, and the resulting control-coefficient matrices are used as compact functional representations. Third, a distance measure is defined according to the differences between fitted trajectories and transformed into an equivalent Euclidean distance through the collocation Gram matrix. Finally, K-means clustering is performed on the transformed coefficient features to identify groups of stocks exhibiting similar dynamic market behavior.

The proposed framework is intended to identify behavioral heterogeneity among stocks within a given sector, rather than to replace existing industry classifications. Its methodological contribution lies in the consistent integration of multi-indicator financial representation, B-spline fitting, and functional-distance-based clustering, rather than in introducing a new spline basis, a new LSPIA algorithm, or a new K-means variant.

The main contributions of this paper are summarized as follows: (i) A multi-indicator representation of stock-market behavior is constructed from three complementary perspectives: periodic returns, risk fluctuations, and trading activity; (ii) a unified B-spline representation of multivariate financial time series is established using LSPIA, providing compact features that preserve the principal temporal patterns of stock-market behavior; (iii) a functional-distance-based feature transformation is developed to establish the equivalence between the distance between fitted B-spline trajectories and the Euclidean distance between transformed control coefficients, enabling K-means clustering in a functionally consistent feature space; and (iv) the proposed framework is applied to stock-behavior clustering within a given sector, and its effectiveness is examined through clustering validity, indicator ablation, parameter sensitivity, and economic interpretation.

The remainder of this paper is organized as follows. Section 2 describes the data and financial indicator construction. Section 3 presents the B-spline fitting model and LSPIA-based representation of multivariate time series. Section 4 develops the functional distance and the corresponding K-means clustering framework. Section 5 reports the empirical results and evaluates the clustering performance and robustness. Finally, Section 6 concludes the paper.

2. Sample selection, data preprocessing, and indicator construction

2.1. Sample selection and data preprocessing

This paper studies A-share stocks from a selected sector. The raw data include the trading date, closing price, highest price, lowest price, trading volume, turnover rate, and trading amount. To ensure data quality and cross-stock comparability, the following preprocessing procedures are performed.

(1) Stocks with insufficient listing histories, prolonged trading suspensions, or missing key market

fields exceeding a prespecified threshold are excluded to ensure sufficient valid observations for subsequent fitting and clustering.

(2) A common price-adjustment convention is applied to closing prices. The highest and lowest prices are maintained under the same convention as the closing price to ensure consistency when checking market-data fields.

(3) Nontrading days are not interpolated. Individual invalid trading days caused by suspension or missing observations are removed according to a unified data-processing rule. After the financial indicators have been constructed, the common valid trading days shared by all sample stocks are extracted.

(4) All stocks are aligned to the common trading-day set, and the resulting financial indicators are standardized using a unified full-sample procedure.

These procedures mitigate the influence of missing observations, inconsistent price conventions, and erroneous records while preserving genuine market fluctuations.

To ensure comparability in subsequent B-spline fitting and clustering, all stocks share the same trading-day window, time parameterization, knot vector, spline degree, and number of control points. Their fitted representations are therefore defined with respect to a common B-spline basis.

2.2. Construction of financial indicators

Before introducing the financial indicators and the subsequent modeling framework, the principal mathematical notation used throughout the paper is summarized in Table 1 for ease of reference.

Table 1. Principal mathematical notation used throughout the paper.

Symbol	Description	Symbol	Description
N	Number of stocks	T	Number of common valid trading days
d	Number of financial indicators	M	Number of B-spline control points
$\mathbf{X}_i$	Standardized multi-indicator matrix of stock i	$\mathbf{A}$	B-spline collocation matrix
$\mathbf{C}_i$	B-spline control-coefficient matrix of stock i	$\mathbf{C}_i^*$	Final control-coefficient matrix obtained by LSPIA
$\widehat{\mathbf{X}}_i$	Fitted multi-indicator B-spline series of stock i	$\mathbf{G}_A$	Gram matrix
d_{ij}	Functional distance between stocks i and j	$\boldsymbol{\phi}_i$	Transformed functional feature vector of stock i
$\boldsymbol{\Phi}$	Transformed feature matrix of all stocks	κ	Number of clusters

For stock i on trading day t, the closing price, trading volume, and raw turnover-rate record are denoted by

$$P_{i,t}^c, \quad V_{i,t}, \quad Q_{i,t},$$

respectively.

Because stocks differ substantially in price level, shares in circulation, and transaction scale, directly using the absolute price, trading volume, or trading amount may cause clustering results to be dominated by firm-size characteristics. To characterize stock-market behavior more appropriately, this paper does not use the closing-price level directly. Instead, four financial indicators are constructed from three perspectives: periodic return, risk exposure, and trading activity.

Specifically, cumulative log return and rolling volatility capture price movements and risk, respectively, while the average rate of change in trading volume and turnover rate reflect dynamic changes and relative levels of trading activity. A common 20-day window is adopted for the rolling indicators to characterize stock behavior over an approximately monthly horizon.

(1) **Cumulative log return**

The daily log return of stock i on trading day t is first defined as

$$r_{i,t} = \ln\left(\frac{P^c_{i,t}}{P^c_{i,t-1}}\right).$$

The daily log return is used to calculate the rolling volatility defined below, but does not enter the B-spline fitting or clustering analysis directly. To characterize periodic price movements, the cumulative log return over a rolling window of length h is defined as

$$\mathbf{R}^{(h)}_{i,t} = \ln\left(\frac{P^c_{i,t}}{P^c_{i,t-h}}\right) = \sum_{s=t-h+1}^{t} r_{i,s}.$$

With $h = 20$, this indicator captures approximately monthly price movements while reducing the influence of individual daily fluctuations on B-spline fitting and stock-distance calculation.

(2) **Average rate of change in trading volume**

Trading volume reflects the intensity of market participation and provides complementary information about investors' trading behavior [22]. The one-day logarithmic rate of change in trading volume is first defined as

$$g^V_{i,t} = \ln\left(\frac{V_{i,t} + \varepsilon^V_i}{V_{i,t-1} + \varepsilon^V_i}\right), \qquad \varepsilon^V_i > 0.$$

Here, ε^V_i is a positive constant introduced to avoid taking the logarithm of zero.

The average logarithmic rate of change in trading volume over a window of length h is then defined as

$$\Delta v^{(h)}_{i,t} = \frac{1}{h}\sum_{s=t-h+1}^{t} g^V_{i,s} = \frac{1}{h}\ln\left(\frac{V_{i,t} + \varepsilon^V_i}{V_{i,t-h} + \varepsilon^V_i}\right).$$

With $h = 20$, this indicator measures the average logarithmic growth or contraction of trading volume over approximately one trading month.

(3) **Turnover rate**

If the turnover rate $Q_{i,t}$ provided by the data source is recorded as a percentage, it is converted to decimal form:

$$\tau_{i,t} = \frac{Q_{i,t}}{100}.$$

No conversion is required if the raw turnover rate is already expressed in decimal form.

The turnover rate measures trading intensity relative to shares in circulation, providing information on trading activity that is more comparable across stocks than absolute trading volume [23].

(4) **Rolling volatility**

To characterize short-term risk exposure, this paper uses the sample standard deviation of the daily log return $r_{i,t}$ over a rolling window of length h [24]:

$$\sigma^{(h)}_{i,t} = \sqrt{\frac{1}{h-1}\sum_{s=t-h+1}^{t}\left(r_{i,s} - \bar{r}^{(h)}_{i,t}\right)^2} \text{ with } \bar{r}^{(h)}_{i,t} = \frac{1}{h}\sum_{s=t-h+1}^{t} r_{i,s}.$$

In this study, $h = 20$ is used to characterize approximately monthly fluctuations in daily returns.

In summary, the raw financial-indicator vector of stock i on trading day t is defined as

$$z_{i,t} = \left(\mathbf{R}_{i,t}^{(20)}, \Delta v_{i,t}^{(20)}, \tau_{i,t}, \sigma_{i,t}^{(20)}\right)^{\top} \in \mathbb{R}^d, \qquad d = 4. \tag{2.1}$$

Because the rolling indicators require historical observations, the first 20 trading days cannot provide complete values for all four indicators. Additional missing or invalid records may also occur for individual stocks.

Let $\mathcal{T}_i$ denote the set of trading days on which stock i has complete and valid observations for all four indicators. The common set of valid trading days for all stocks is

$$\mathcal{T} = \bigcap_{i=1}^{N} \mathcal{T}_i.$$

The trading days in $\mathcal{T}$ are reordered chronologically and indexed as $t = 1, 2, \ldots, T$, where $T = |\mathcal{T}|$ denotes the number of valid trading days shared by all sample stocks. The subsequent standardization of financial indicators, B-spline fitting, functional-distance calculation, and stock clustering are all based on this common set of trading days.

The complementarity of the four indicators is further examined through correlation analysis and one-indicator-at-a-time ablation experiments.

2.3. Feature standardization and input matrices

The four financial indicators differ substantially in unit, range, and fluctuation magnitude. If B-spline fitting and distance calculation are conducted directly, indicators with larger numerical scales may dominate the clustering process. To improve comparability among indicators, a full-sample standardization procedure is used.

Let $z_{i,t}^{(\ell)}$ denote the ℓth component of $z_{i,t}$, where $\ell = 1, 2, \ldots, d$ and $d = 4$ in this paper. Over all N stocks and T common trading days, the sample mean and sample standard deviation of indicator ℓ are defined as

$$\mu_\ell = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{i,t}^{(\ell)} \quad \text{and} \quad s_\ell = \sqrt{\frac{1}{NT-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(z_{i,t}^{(\ell)} - \mu_\ell\right)^2}.$$

The standardized indicator is then

$$x_{i,t}^{(\ell)} = \frac{z_{i,t}^{(\ell)} - \mu_\ell}{s_\ell}.$$

After standardization, the financial indicators are transformed to comparable numerical scales, thereby preventing undue influence of differences in numerical range and fluctuation magnitude on LSPIA fitting and stock-distance calculation. This procedure performs a location-and-scale transformation of the input data and does not change the mathematical form of the subsequent LSPIA least-squares model.

This paper adopts full-sample standardization, in which the mean and standard deviation of each indicator are calculated jointly over all stocks and common trading days. Unlike stock-specific standardization, which removes individual differences in indicator levels, full-sample standardization preserves both cross-sectional heterogeneity and temporal variation while placing the indicators on

comparable scales. This choice is consistent with the objective of identifying stocks with similar market behavior in terms of both dynamic patterns and persistent indicator characteristics. By contrast, time-specific cross-sectional standardization emphasizes relative positions among stocks but removes market-wide temporal shifts. The robustness of the clustering results under alternative standardization schemes is further examined in the empirical analysis.

Finally, the standardized multi-indicator input matrix for stock i is denoted by

$$\mathbf{X}_i = \begin{bmatrix} x_{i,1}^{(1)} & x_{i,1}^{(2)} & \cdots & x_{i,1}^{(d)} \\ x_{i,2}^{(1)} & x_{i,2}^{(2)} & \cdots & x_{i,2}^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,T}^{(1)} & x_{i,T}^{(2)} & \cdots & x_{i,T}^{(d)} \end{bmatrix} \in \mathbb{R}^{T\times d}. \tag{2.2}$$

The collection $\{\mathbf{X}_i\}_{i=1}^N$ constitutes the standardized input data for subsequent B-spline fitting.

3. B-spline-LSPIA representation of multi-indicator time series

After constructing and standardizing the financial indicators, this paper uses a common cubic B-spline basis to fit the multi-indicator time series $\mathbf{X}_i$ of each stock by least squares. Solving for the corresponding control-coefficient matrix using LSPIA yields control coefficients that provide a compact functional representation of stock-market behavior for subsequent clustering.

3.1. Unified time parameters and the B-spline collocation matrix

To ensure that the B-spline control coefficients of different stocks have the same meaning and are mutually comparable, all stocks use the same time parameters, knot vector, B-spline degree, and number of control points.

All common trading days are linearly mapped to the interval [0, 1], and the common time parameters are defined as

$$\tau_t = \frac{t-1}{T-1}, \qquad t = 1, 2, \ldots, T. \tag{3.1}$$

Clearly, $0 = \tau_1 < \tau_2 < \cdots < \tau_T = 1$. Let M denote the number of control points of the cubic B-spline, with $M \ll T$ in general. A common open uniform knot vector is adopted:

$$\mathbf{u} = (u_0, u_1, \ldots, u_{M+3}), \tag{3.2}$$

where

$$u_r = \begin{cases} 0, & 0 \le r \le 3, \\ \dfrac{r-3}{M-3}, & 4 \le r \le M-1, \\ 1, & M \le r \le M+3. \end{cases}$$

Let $\{N_m(\tau)\}_{m=0}^{M-1}$ be the cubic B-spline basis functions defined on the knot vector $\mathbf{u}$. Then the common B-spline collocation matrix constructed from the time parameters of all trading days is

$$\mathbf{A} = \begin{bmatrix} N_0(\tau_1) & N_1(\tau_1) & \cdots & N_{M-1}(\tau_1) \\ N_0(\tau_2) & N_1(\tau_2) & \cdots & N_{M-1}(\tau_2) \\ \vdots & \vdots & \ddots & \vdots \\ N_0(\tau_T) & N_1(\tau_T) & \cdots & N_{M-1}(\tau_T) \end{bmatrix} \in \mathbb{R}^{T\times M}. \tag{3.3}$$

Since the time parameters, spline degree, knot vector, and number of control points are identical for all stocks, the collocation matrix $\mathbf{A}$ is constructed only once and shared throughout the fitting process. Consequently, the control-coefficient matrices are expressed in a common B-spline coordinate system, providing a consistent basis for subsequent functional-distance calculation and clustering.

3.2. Multi-indicator B-spline least-squares fitting model

For stock i, let its B-spline control-coefficient matrix be

$$\mathbf{C}_i = \begin{bmatrix} c_{i,1}^{(1)} & c_{i,1}^{(2)} & \cdots & c_{i,1}^{(d)} \\ c_{i,2}^{(1)} & c_{i,2}^{(2)} & \cdots & c_{i,2}^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ c_{i,M}^{(1)} & c_{i,M}^{(2)} & \cdots & c_{i,M}^{(d)} \end{bmatrix} \in \mathbb{R}^{M\times d}.$$

The ℓth column,

$$\mathbf{c}_i^{(\ell)} = \left(c_{i,1}^{(\ell)}, c_{i,2}^{(\ell)}, \ldots, c_{i,M}^{(\ell)}\right)^{\top} \in \mathbb{R}^{M},$$

contains the B-spline control coefficients associated with the ℓth financial indicator of stock i. The control coefficients are determined using the least-squares criterion:

$$\min_{\mathbf{C}_i\in\mathbb{R}^{M\times d}} f_i(\mathbf{C}_i) := \frac{1}{2}\left\|\mathbf{X}_i - \mathbf{A}\mathbf{C}_i\right\|_F^2. \tag{3.4}$$

Under the common collocation matrix $\mathbf{A}$, the multi-indicator B-spline fitted matrix of stock i is

$$\widehat{\mathbf{X}}_i = \mathbf{A}\mathbf{C}_i \in \mathbb{R}^{T\times d}. \tag{3.5}$$

The ℓth column of $\widehat{\mathbf{X}}_i$ is the B-spline fit to the time series of the ℓth financial indicator, and $\mathbf{C}_i$ is the unknown control-coefficient matrix.

Model (3.4) is equivalent to fitting each financial indicator independently under the same B-spline basis. Thus, the indicators share a common temporal basis while retaining their respective control coefficients. The resulting matrix $\mathbf{C}_i$ provides a compact functional representation of the multivariate time series $\mathbf{X}_i$. The objective is to approximate and compress the temporal patterns of stock-market behavior, rather than to explicitly model causal relationships among the financial indicators.

3.3. LSPIA iteration for the control-coefficient matrix

To avoid solving the normal equations directly, LSPIA is used to solve (3.4) iteratively. The initial control coefficients are obtained by approximately uniform sampling of the standardized input matrix $\mathbf{X}_i$, providing an initial representation of its temporal pattern.

Define the initial sampling positions by

$$s_m = 1 + \operatorname{round}\left(\frac{(m-1)(T-1)}{M-1}\right), \qquad m = 1, 2, \ldots, M.$$

The initial control-coefficient matrix is then

$$\mathbf{C}_i^{(0)}(m,:) = \mathbf{X}_i(s_m,:), \qquad m = 1, 2, \ldots, M, \tag{3.6}$$

where $\mathbf{X}_i(s_m, :)$ denotes row s_m of $\mathbf{X}_i$.

Let $\mathbf{C}_i^{(k)}$ be the control-coefficient matrix at iteration k. Then the fitted matrix and residual matrix are

$$\widehat{\mathbf{X}}_i^{(k)} = \mathbf{A}\mathbf{C}_i^{(k)} \tag{3.7}$$

and

$$\mathbf{R}_i^{(k)} = \mathbf{X}_i - \widehat{\mathbf{X}}_i^{(k)} = \mathbf{X}_i - \mathbf{A}\mathbf{C}_i^{(k)},$$

respectively. The matrix form of the LSPIA iteration is

$$\mathbf{C}_i^{(k+1)} = \mathbf{C}_i^{(k)} + \mu\mathbf{A}^\top\mathbf{R}_i^{(k)} = \mathbf{C}_i^{(k)} + \mu\mathbf{A}^\top\left(\mathbf{X}_i - \mathbf{A}\mathbf{C}_i^{(k)}\right), \tag{3.8}$$

where $\mu > 0$ is the iterative step size.

Define

$$\mathbf{G}_A = \mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{M\times M}.$$

As stated in [14, 16], when the step size satisfies

$$0 < \mu < \frac{2}{\lambda_{\max}(\mathbf{G}_A)}, \tag{3.9}$$

the LSPIA iteration converges to a least-squares solution of (3.4). Moreover, for a fixed scalar step size, the asymptotic convergence factor is minimized by

$$\mu = \frac{2}{\lambda_{\min}(\mathbf{G}_A) + \lambda_{\max}(\mathbf{G}_A)}.$$

In the numerical implementation, the iteration is terminated when the change between two successive control-coefficient matrices satisfies

$$\left\|\mathbf{C}_i^{(k+1)} - \mathbf{C}_i^{(k)}\right\|_F < \text{tol}, \tag{3.10}$$

where tol is a prescribed convergence tolerance. If the stopping criterion is not satisfied within the maximum number $k_{\max}$ of iterations, the iteration is terminated and its convergence status is recorded.

Let the control-coefficient matrix at termination be

$$\mathbf{C}_i^* = \mathbf{C}_i^{(k_i^{\text{end}})}, \tag{3.11}$$

where k_i^{end} denotes the actual number of LSPIA iterations performed for stock i.

The final matrix $\mathbf{C}_i^* \in \mathbb{R}^{M\times d}$ represents the original $T \times d$ time series using only $M \times d$ control coefficients. Since $M \ll T$, this representation reduces the temporal dimensionality while retaining the main fitted patterns of the financial indicators. All subsequent clustering analyses are based on the collection of final control-coefficient matrices $\{\mathbf{C}_i^*\}_{i=1}^N$. The functional distance between stocks is constructed from their corresponding fitted B-spline trajectories, as developed in the next section.

4. Stock-behavior clustering based on control-coefficient features

After LSPIA iteration, the standardized multi-indicator time series $\mathbf{X}_i$ of stock i is compressed into the control-coefficient matrix $\mathbf{C}_i^*$. This matrix can be regarded as a low-dimensional functional representation of the stock's multi-indicator market-behavior series. Compared with clustering directly from the raw time series, the control-coefficient representation reduces dimensionality, weakens the influence of high-frequency noise, and preserves the overall dynamic structures of returns, risk, and trading activity across stocks.

4.1. Functional distance based on fitted B-spline curves

Directly using the Frobenius distance $\left\|\mathbf{C}_i^* - \mathbf{C}_j^*\right\|_F$ between control-coefficient matrices implicitly assumes that all coefficients contribute equally to the fitted results. Because B-spline basis functions have local support and overlap, however, the influence of different control coefficients on the fitted series also depends on the collocation matrix $\mathbf{A}$. Therefore, stock similarity is measured through the corresponding fitted B-spline trajectories rather than directly through the raw control coefficients.

According to (3.5) and (3.11), the final fitted matrix of stock i is

$$\widehat{\mathbf{X}}_i = \mathbf{A}\mathbf{C}_i^*.$$

The squared discrete functional distance between stocks i and j is defined as

$$d_{ij}^2 = \left\|\widehat{\mathbf{X}}_i - \widehat{\mathbf{X}}_j\right\|_F^2 = \left\|\mathbf{A}\mathbf{C}_i^* - \mathbf{A}\mathbf{C}_j^*\right\|_F^2. \tag{4.1}$$

This distance measures the aggregate squared differences between the fitted values of all financial indicators over the common trading days.

Let

$$\Delta\mathbf{C}_{ij} = \mathbf{C}_i^* - \mathbf{C}_j^*.$$

Then, using the Gram matrix $\mathbf{G}_A = \mathbf{A}^\top\mathbf{A}$ defined in Section 3, (4.1) can be rewritten as

$$d_{ij}^2 = \left\|\mathbf{A}\Delta\mathbf{C}_{ij}\right\|_F^2 = \mathrm{tr}\left(\Delta\mathbf{C}_{ij}^\top\mathbf{A}^\top\mathbf{A}\Delta\mathbf{C}_{ij}\right) = \mathrm{tr}\left(\Delta\mathbf{C}_{ij}^\top\mathbf{G}_A\Delta\mathbf{C}_{ij}\right). \tag{4.2}$$

Remark 4.1. *The distance in* (4.1) *is defined on the common discrete trading-day grid rather than through continuous integration over the parameter domain. It incorporates the local support and overlap of the B-spline basis through the Gram matrix* $\mathbf{G}_A$*. Consequently, it is generally different from the unweighted Frobenius distance between the original control-coefficient matrices.*

The stock-distance matrix constructed from all pairwise distances is

$$\mathbf{D}_{\text{stock}} = \left[d_{ij}\right]_{N\times N}. \tag{4.3}$$

By construction, $\mathbf{D}_{\text{stock}}$ is symmetric, has zero diagonal entries, and contains nonnegative distances. When $\mathbf{A}$ has full column rank, the induced distance is a metric on the control-coefficient representations.

4.2. Equivalent vectorized features of the control coefficients

To apply standard Euclidean clustering algorithms, the functional distance is transformed into an equivalent distance between vectorized coefficient features.

Let $\mathbf{G}_A^{1/2}$ be the principal square root of the symmetric positive-semidefinite matrix $\mathbf{G}_A$, satisfying

$$\left(\mathbf{G}_A^{1/2}\right)^\top\mathbf{G}_A^{1/2} = \mathbf{G}_A.$$

Define the vectorized control-coefficient feature of stock i by

$$\boldsymbol{\phi}_i = \mathrm{vec}\left(\mathbf{G}_A^{1/2}\mathbf{C}_i^*\right) \in \mathbb{R}^{Md}. \tag{4.4}$$

Then, using the identity $\|\operatorname{vec}(\mathbf{B})\|_2 = \|\mathbf{B}\|_F$, the squared Euclidean distance between two transformed features satisfies

$$\left\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\right\|_2^2 = \left\|\mathbf{G}_A^{1/2}\Delta\mathbf{C}_{ij}\right\|_F^2 = \operatorname{tr}\left(\Delta\mathbf{C}_{ij}^{\top}\mathbf{G}_A\Delta\mathbf{C}_{ij}\right) = d_{ij}^2. \tag{4.5}$$

Combining (4.2) and (4.5) yields

$$d_{ij} = \left\|\mathbf{A}\mathbf{C}_i^* - \mathbf{A}\mathbf{C}_j^*\right\|_F = \left\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\right\|_2. \tag{4.6}$$

Equation (4.6) establishes an exact equivalence between the discrete functional distance between the fitted B-spline trajectories and the Euclidean distance between their transformed control-coefficient features.

Remark 4.2. *The equivalence in* (4.6) *is an exact reparameterization: Applying K-means to the reconstructed matrices* $\widehat{\mathbf{X}}_i = \mathbf{A}\mathbf{C}_i^*$ *(after vectorization) and applying it to* $\boldsymbol{\phi}_i$ *give the same squared pairwise distances and the same K-means objective for corresponding partitions. Thus, no improvement in clustering outcomes should be attributed to choosing one of these two* equivalent *representations. The distinction is instead between the fitted-series distance and the unweighted Euclidean distance between the original coefficients, which need not coincide because* $\mathbf{A}^{\top}\mathbf{A}$ *is generally not the identity.*

For an illustrative two-coefficient example with one indicator, consider the common collocation matrix and two coefficient differences

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{bmatrix}, \qquad \Delta\mathbf{C}^{(a)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \Delta\mathbf{C}^{(b)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Both have the same raw coefficient distance, $\|\Delta\mathbf{C}^{(a)}\|_2 = \|\Delta\mathbf{C}^{(b)}\|_2 = \sqrt{2}$*. However, their fitted-series distances differ:*

$$\|\mathbf{A}\Delta\mathbf{C}^{(a)}\|_2 = \sqrt{3}, \qquad \|\mathbf{A}\Delta\mathbf{C}^{(b)}\|_2 = \sqrt{2}.$$

The Gram-transformed features reproduce these values exactly because $\|\mathbf{G}_A^{1/2}\Delta\mathbf{C}\|_2 = \|\mathbf{A}\Delta\mathbf{C}\|_2$*. Hence, the transformation distinguishes coefficient changes that have different effects on the reconstructed trajectories, whereas raw coefficient distances treat them as equal.*

There is also a practical representation advantage once a common basis and its Gram factor have been computed. A stock is represented by Md *transformed coefficients rather than* Td *fitted values (*400 *rather than* 6028 *in this study). Evaluating one pairwise Euclidean distance then requires* $O(Md)$ *operations instead of* $O(Td)$ *operations on explicitly reconstructed trajectories, excluding the one-time fitting and transformation costs. This is a feature-storage and subsequent-distance-evaluation advantage, not evidence of a smaller end-to-end runtime than every alternative.*

Arranging the vectorized features of all stocks by rows gives the feature matrix

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N]^{\top} \in \mathbb{R}^{N\times Md}. \tag{4.7}$$

Thus, the original N multivariate time series, each containing $T \times d$ observations, are represented by an $N \times Md$ feature matrix. Since $M \ll T$, the dimensionality involved in subsequent clustering is substantially reduced.

The transformation preserves the pairwise distances between fitted trajectories exactly, although the fitted trajectories themselves are approximations of the original standardized observations.

In practical computation, any matrix factor $\mathbf{L} \in \mathbb{R}^{M\times M}$ satisfying $\mathbf{L}^{\top}\mathbf{L} = \mathbf{G}_A$ can replace $\mathbf{G}_A^{1/2}$ without changing the pairwise distances. For example, a Cholesky factorization can be used when $\mathbf{G}_A$ is positive definite.

4.3. K-means clustering

The transformed feature matrix $\mathbf{\Phi}$ provides an explicit Euclidean representation of stock-market behavior, allowing standard K-means to be applied without modifying its clustering objective [25].

For a prescribed number of clusters κ, K-means partitions the N stocks into κ mutually disjoint, nonempty clusters $C_1, C_2, \ldots, C_\kappa$ by solving

$$\min_{C_1,\ldots,C_\kappa} \sum_{r=1}^{\kappa} \sum_{i\in C_r} \left\|\boldsymbol{\phi}_i - \mathbf{m}_r\right\|_2^2, \tag{4.8}$$

where

$$\mathbf{m}_r = \frac{1}{|C_r|} \sum_{i\in C_r} \boldsymbol{\phi}_i$$

is the center of cluster r in the transformed feature space. The distance equivalence established above also provides a functional interpretation of the K-means objective.

Define the mean control-coefficient matrix of cluster r as

$$\overline{\mathbf{C}}_r = \frac{1}{|C_r|} \sum_{i\in C_r} \mathbf{C}_i^*.$$

Since the feature transformation is linear, the cluster center satisfies $\mathbf{m}_r = \mathrm{vec}\left(\mathbf{G}_A^{1/2}\overline{\mathbf{C}}_r\right)$. Consequently,

$$\sum_{r=1}^{\kappa} \sum_{i\in C_r} \|\boldsymbol{\phi}_i - \mathbf{m}_r\|_2^2 = \sum_{r=1}^{\kappa} \sum_{i\in C_r} \left\|\mathbf{G}_A^{1/2}(\mathbf{C}_i^* - \overline{\mathbf{C}}_r)\right\|_F^2 = \sum_{r=1}^{\kappa} \sum_{i\in C_r} \left\|\mathbf{A}(\mathbf{C}_i^* - \overline{\mathbf{C}}_r)\right\|_F^2. \tag{4.9}$$

Thus, minimizing the Euclidean K-means objective in the transformed feature space is equivalent to minimizing the within-cluster squared differences between fitted B-spline trajectories and their corresponding cluster-mean trajectories. This equivalence ensures that the clustering objective is consistent with the functional distance defined in (4.1).

By alternating between sample assignment and cluster-center updating, K-means produces a label for each stock:

$$\ell_i^{\mathrm{km}} \in \{1, 2, \ldots, \kappa\}, \qquad i = 1, 2, \ldots, N.$$

Since K-means favors relatively compact clusters in Euclidean space, its suitability for the present data is assessed through clustering validity, cluster sizes, and sensitivity analyses.

4.4. Evaluation of clustering results and selection of the number of clusters

For a given number M of control points and number κ of clusters, K-means is performed on the control-coefficient feature matrix $\mathbf{\Phi}$. The clustering results are evaluated from the perspectives of internal clustering quality, cluster-size distribution, stability, and economic interpretability.

(1) Internal clustering indices

The silhouette coefficient, Calinski-Harabasz (CH) index, and Davies-Bouldin (DB) index are used to evaluate within-cluster compactness and between-cluster separation [26–28]. Higher silhouette and CH values and lower DB values generally indicate more favorable internal clustering structures. The silhouette coefficient is computed using either Euclidean distances in $\mathbf{\Phi}$ or the equivalent functional-distance matrix $\mathbf{D}_{\text{stock}}$, whereas the CH and DB indices are calculated from $\mathbf{\Phi}$. The number of clusters is examined by comparing these indices across candidate values of κ, together with cluster-size distributions and the economic characteristics of the resulting groups.

(2) Economic interpretability

The identified clusters are characterized using the four original financial indicators: 20-day cumulative log return, 20-day average rate of change in trading volume, turnover rate, and 20-day rolling volatility. Differences in their indicator levels and temporal patterns are examined to determine whether the resulting groups exhibit distinguishable market-behavior characteristics. In addition, indicator-ablation experiments and parameter-sensitivity analyses are conducted to assess the contributions of individual indicators and the robustness of the clustering results.

The complete procedure, including data preprocessing, LSPIA-B-spline fitting, functional-distance construction, feature transformation, K-means clustering, and evaluation, is summarized in Algorithm 1.

5. Empirical analysis

In this section, the empirical analysis is conducted using A-share stocks from the pharmaceutical sector. Following the preprocessing and indicator-construction procedures described in Section 2, the four financial indicators are calculated and aligned over the common valid trading days.

All experiments are conducted in MATLAB. The window length for the rolling indicators is set to $h = 20$. The baseline number of LSPIA control points is $M = 100$, while $M = 80, 90, 110, 120$ are considered in the parameter-sensitivity analysis. The maximum number of iterations is $k_{\max} = 5000$, and the stopping tolerance is $\text{tol} = 10^{-8}$. Candidate cluster numbers are $\kappa = 2, 3, 4, 5, 6, 7, 8$. To reduce the influence of random initialization, K-means++ is used to select the initial cluster centers. For each candidate number of clusters, K-means is run independently 30 times, and the result with the minimum within-cluster sum of squares is retained. The random-number seed is fixed at 2026 to ensure reproducibility.

The fitting quality of LSPIA is evaluated using the root mean squared error (RMSE) and mean absolute error (MAE). Their averages over all stocks are defined as

$$\text{RMSE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{Td}\left\|\mathbf{X}_i - \mathbf{A}\mathbf{C}_i^*\right\|_F^2}$$

and

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Td}\sum_{t=1}^{T}\sum_{\ell=1}^{d}\left|x_{i,t}^{(\ell)} - (\mathbf{A}\mathbf{C}_i^*)_{t,\ell}\right|.$$

RMSE emphasizes larger fitting deviations, whereas MAE measures the average absolute deviation. Both metrics are calculated from standardized indicators.

Algorithm 1: LSPIA-B-spline-based stock-behavior clustering

Input: Stock-market data; number of control points M; indicator-window length h; LSPIA parameters $k_{\max}$ and tol; candidate cluster set $\mathcal{K}$; number of K-means repetitions n_{rep}.

Output: Final cluster labels $\{\ell_i^*\}_{i=1}^N$; control-coefficient matrices $\{\mathbf{C}_i^*\}_{i=1}^N$; selected cluster number κ^*; clustering-evaluation results.

Stage 1: Financial indicator construction.

Extract closing prices, trading volumes, and turnover rates; screen stocks and align valid trading days;

Construct the four financial indicators using window length h and standardize them to obtain $\mathbf{X}_i \in \mathbb{R}^{T\times 4}$, $i = 1, \ldots, N$;

Stage 2: LSPIA-based feature representation.

Construct the common time parameters, cubic B-spline knot vector, and collocation matrix $\mathbf{A}$ according to (3.1)–(3.3); compute $\mathbf{G}_A = \mathbf{A}^\top\mathbf{A}$ and the step size μ;

for $i = 1, \ldots, N$ **do**

    Initialize $\mathbf{C}_i^{(0)}$ according to (3.6);

    for $k = 0, \ldots, k_{\max} - 1$ **do**

        Update $\mathbf{C}_i^{(k+1)}$ using (3.8);

        if (3.10) *is satisfied* **then**

            break;

    Save $\mathbf{C}_i^*$ according to (3.11);

Compute $\mathbf{D}_{\text{stock}}$ using (4.3); construct $\boldsymbol{\phi}_i$ and $\boldsymbol{\Phi}$ according to (4.4) and (4.7);

Stage 3: K-means clustering and evaluation.

for $\kappa \in \mathcal{K}$ **do**

    Run K-means on $\boldsymbol{\Phi}$ independently n_{rep} times and retain the solution with the minimum within-cluster sum of squares;

    Compute the silhouette, CH, and DB indices; record the cluster sizes;

Stage 4: Final cluster selection.

Select κ^* by jointly considering the clustering indices, cluster sizes, and financial-indicator characteristics;

Return $\{\ell_i^*\}_{i=1}^N$, $\{\mathbf{C}_i^*\}_{i=1}^N$, and κ^*;

The temporal compression ratio is defined as $\rho = 1 - M/T$. It measures the reduction in the number of temporal coefficients per indicator and does not represent the actual memory usage of the complete algorithm.

To evaluate K-means under different numbers of clusters, the silhouette coefficient, CH index, and DB index are used together with the minimum and maximum cluster sizes.

The adjusted Rand index (ARI) is used to assess agreement between clustering partitions under different experimental settings [29]. An ARI of 1 indicates identical partitions, whereas a value close to 0 indicates agreement comparable to that expected by chance. ARI is invariant to cluster-label permutations and does not require external ground-truth labels.

5.1. Data description and indicator characteristics

After sample screening and alignment of common trading days, $N = 113$ stocks and $T = 1507$ common valid trading days are retained. The sample period extends from November 12, 2018, to December 26, 2025. Table 2 reports the descriptive statistics of the four indicators before standardization.

As shown in Table 2, the cumulative log return and average volume-change rate have means close to zero but exhibit considerable dispersion, reflecting heterogeneous price movements and trading-volume changes across the sample. The turnover rate and rolling volatility are nonnegative, with means of 0.0220 and 0.0243, respectively. Their maxima, 0.5020 and 0.1467, are substantially higher than their corresponding medians, indicating pronounced upper-tail observations and heterogeneity in trading activity and risk fluctuations. The indicators also differ considerably in scale, as reflected by the standard deviations of cumulative log return (0.1218) and rolling volatility (0.0122). This motivates the unified standardization procedure adopted before B-spline fitting and clustering.

Table 2. Descriptive statistics of the four financial indicators.

Indicator	Mean	Standard deviation	Median	Minimum	Maximum
20-day cumulative log return	0.0014	0.1218	-0.0033	-0.9282	1.8610
20-day average rate of change in trading volume	0.0003	0.0395	-0.0016	-0.2803	0.2855
Turnover rate	0.0220	0.0305	0.0127	0.0000	0.5020
20-day rolling volatility	0.0243	0.0122	0.0215	0.0025	0.1467

Table 3 reports the Pearson correlations among the four indicators. We can see that all pairwise correlations are positive, ranging from 0.1859 to 0.5366. The strongest correlation occurs between turnover rate and rolling volatility, suggesting an association between trading activity and risk fluctuations. The remaining correlations indicate that the indicators contain overlapping but nonidentical linear information. Their effects on clustering are further examined through indicator-ablation experiments.

Table 3. Pearson correlations among the four financial indicators.

Indicator	20-day cumulative return	20-day average volume change	Turnover rate	20-day rolling volatility
20-day cumulative return	1.0000	0.4414	0.3720	0.1980
20-day average volume change	0.4414	1.0000	0.4017	0.1859
Turnover rate	0.3720	0.4017	1.0000	0.5366
20-day rolling volatility	0.1980	0.1859	0.5366	1.0000

5.2. LSPIA fitting performance and control-point selection

The number M of control points determines both the representational capacity of the B-spline model and the dimensionality of the control-coefficient features. Increasing M generally improves fitting flexibility but reduces compression and increases feature dimensionality. All 113 stocks satisfy the prescribed stopping criterion under every tested value of M. The fitting results are reported in Table 4.

As M increases from 80 to 120, the average iteration count remains relatively stable, ranging from 231.2212 to 235.2743, while the average RMSE decreases monotonically from 0.5522 to 0.4640 and the compression ratio decreases from 94.6914% to 92.0372%. The improvement in fitting accuracy becomes less pronounced as M increases.

Table 4. LSPIA fitting results under different numbers of control points.

M	Feature dimension Md	Compression ratio	Average RMSE	Average iterations	Maximum terminal error	Converged stocks
80	320	94.6914%	0.5522	235.2743	9.9947×10^{-9}	113
90	360	94.0279%	0.5205	233.8142	9.9958×10^{-9}	113
100	400	93.3643%	0.4929	232.4071	9.9925×10^{-9}	113
110	440	92.7007%	0.4749	231.3982	9.9986×10^{-9}	113
120	480	92.0372%	0.4640	231.2212	9.9988×10^{-9}	113

The fitting quality of each standardized indicator is further evaluated using RMSE, MAE, and the coefficient of determination R^2. The results for $M = 100$ are reported in Table 5. Rolling volatility achieves the highest fitting accuracy, with an R^2 of 0.9305, followed by cumulative log return and turnover rate, with R^2 values of 0.8022 and 0.7298, respectively. The volume-change indicator has the lowest R^2 (0.5134), indicating that a larger proportion of its variation remains unexplained by the B-spline fit. These results suggest that the indicators exhibit different degrees of temporal smoothness, with the B-spline representation capturing their principal dynamic patterns rather than individual short-term fluctuations.

Table 5. LSPIA fitting quality for the four financial indicators ($M = 100$).

Indicator	RMSE	MAE	R^2
20-day cumulative log return	0.4447	0.3139	0.8022
20-day average rate of change in trading volume	0.6976	0.5331	0.5134
Turnover rate	0.5198	0.2526	0.7298
20-day rolling volatility	0.2637	0.1825	0.9305

At the baseline setting $M = 100$, the average LSPIA iteration count across the 113 stocks is 232.4071, and the maximum terminal control-coefficient correction is 9.9925×10^{-9}. Each 1507×4 multi-indicator time series is compressed into a 100×4 control-coefficient matrix, achieving an average RMSE of 0.4929 and a temporal compression ratio of 93.3643%. Therefore, $M = 100$ is provisionally selected as the baseline setting by balancing fitting accuracy, compression, and convergence performance. Its influence on the final clustering results is further examined in Section 5.4.2.

5.3. Clustering results and market-behavior interpretation

Using the provisionally selected number of control points $M = 100$, K-means is applied to the transformed feature matrix $\mathbf{\Phi}$ for candidate cluster numbers $\kappa = 2, 3, \ldots, 8$.

The silhouette coefficient, CH index, DB index, and cluster-size distribution are reported in Table 6 and Figure 1.

At $\kappa = 2$, the silhouette coefficient and CH index reach their highest values among the tested candidates, at 0.1687 and 14.1668, respectively. The two clusters contain 38 and 75 stocks, with neither forming an abnormally small group. Nevertheless, the modest absolute silhouette value indicates substantial overlap and limited separation between the resulting clusters. For $\kappa = 3$ and $\kappa = 4$, both the silhouette coefficient and CH index decrease. Although some larger values of κ yield lower DB indices, every partition with $\kappa \geq 5$ contains a singleton cluster. Considering internal validity and cluster-size distribution jointly, the baseline number of clusters is selected as $\kappa^* = 2$. This choice provides a

descriptive two-group partition within the tested candidate range, rather than evidence of two sharply separated or uniquely determined market structures.

Table 6. K-means clustering results under different numbers of clusters ($M = 100$).

κ	Silhouette	CH	DB	Minimum cluster size	Maximum cluster size
2	0.1687	14.1668	2.7158	38	75
3	0.0672	10.4795	2.9663	22	58
4	0.0733	8.8174	2.8613	7	55
5	0.0788	7.7370	2.3987	1	64
6	0.0401	7.3129	2.3802	1	47
7	0.0165	6.3920	2.6804	1	34
8	0.0131	6.0412	2.4205	1	37

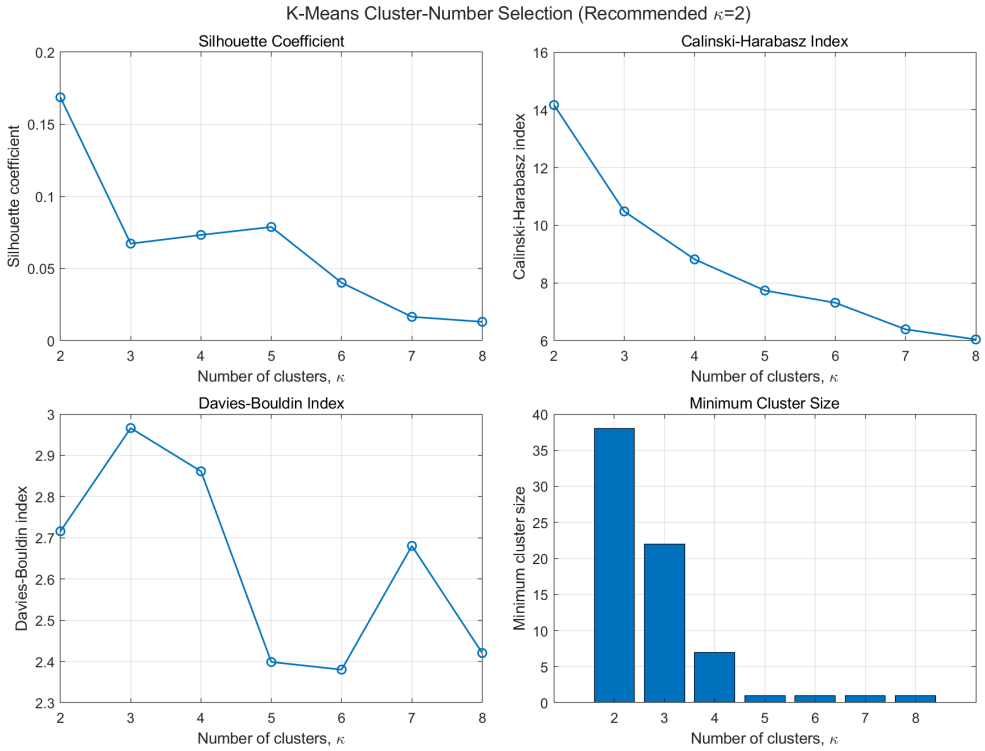


Figure 1. K-means evaluation results under different numbers of clusters.

Principal component analysis (PCA) is applied to the transformed feature matrix $\mathbf{\Phi}$ to visualize the stock distribution in two dimensions. The projection is displayed in Figure 2, with points colored according to their K-means labels. The first two principal components explain 16.2929% and 8.0029% of the total feature variance, respectively, giving a cumulative explained variance of 24.2958%.

Figure 2 shows substantial overlap between the two groups, with limited separation along the first principal-component direction. This observation is consistent with the modest overall silhouette coefficient of 0.1687. Since the first two principal components explain only 24.2958% of the feature variance, the projection provides an auxiliary visualization rather than a complete representation of the original feature-space geometry.

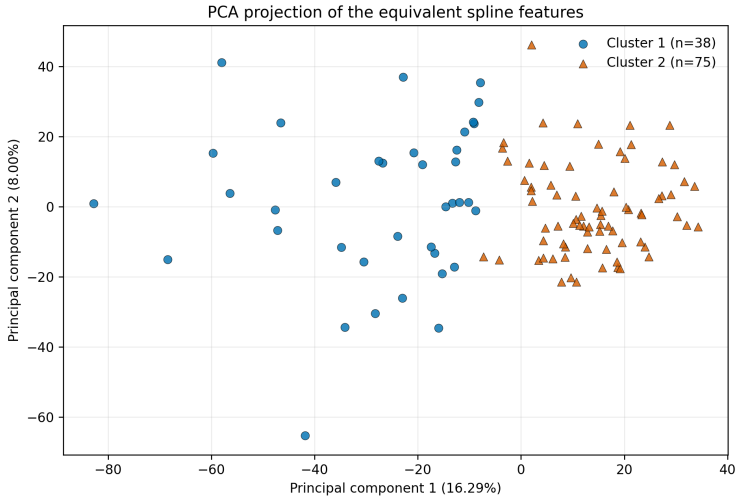


Figure 2. Two-dimensional PCA projection of the control-coefficient features and K-means clustering results.

The average within-cluster functional distances are 98.8488 and 65.2225 for Clusters 1 and 2, respectively, whereas the average between-cluster distance is 92.1562. Cluster 2 is relatively compact, with a smaller within-cluster distance than the between-cluster average. In contrast, Cluster 1 exhibits greater internal heterogeneity, as its within-cluster distance exceeds the average between-cluster distance. The average silhouette values are −0.0749 for Cluster 1 and 0.2921 for Cluster 2. In particular, the negative mean for Cluster 1 indicates that its members, on average, are not more similar to their assigned group than to the alternative group under the adopted distance. Hence the overall silhouette coefficient of 0.1687 does not support a claim of two well-separated behavioral categories. The result is retained as an exploratory, economically interpretable partition, with Cluster 1 being a heterogeneous group rather than a compact cluster.

The economic characteristics of each cluster are examined using the original, unstandardized financial indicators. Let C_r denote cluster r. Its mean value for indicator ℓ is defined as

$$\overline{z}_r^{(\ell)} = \frac{1}{|C_r|T} \sum_{i \in C_r} \sum_{t=1}^{T} z_{i,t}^{(\ell)}.$$

These cluster-level means provide a descriptive basis for interpreting differences in periodic returns, risk fluctuations, and trading activity.

To interpret the resulting stock groups, the four original financial indicators are averaged within each cluster. The results are reported in Table 7.

Table 7. Market-behavior characteristics of the final K-means clusters.

Cluster	Number of stocks	Cumulative return	Volume-change rate	Turnover rate	Rolling volatility	Behavior type
1	38	0.0023	0.0004	0.0345	0.0277	Higher-turnover, higher-volatility group
2	75	0.0009	0.0002	0.0157	0.0226	Lower-turnover, lower-volatility group

Table 7 shows that Cluster 1 has higher mean values for all four indicators. The most pronounced differences occur in turnover rate and rolling volatility. Specifically, the average turnover rate

of Cluster 1 is approximately 2.20 times that of Cluster 2, while its average rolling volatility is approximately 1.23 times that of Cluster 2. These differences suggest that trading intensity and risk fluctuations are important dimensions of the observed behavioral distinction. The mean cumulative log return and volume-change rate are also higher in Cluster 1. However, both indicators have relatively small temporal means, suggesting that their differences are less pronounced in the aggregate cluster profiles.

Accordingly, Cluster 1 is described as a "higher-turnover, higher-volatility" group, whereas Cluster 2 is described as a "lower-turnover, lower-volatility" group. Representative stocks selected by the program include 000153, 000518, 000590, 000788, and 000908 for Cluster 1, and 000504, 000650, 000739, 000913, and 000919 for Cluster 2. These descriptions are relative to the observed sample and do not imply fixed economic categories, future performance, or causal relationships. Cluster labels have no intrinsic ordinal meaning.

The standardized cluster profiles are shown in Figure 3. Cluster 1 exhibits positive mean standardized values for all four indicators, particularly turnover rate and rolling volatility, whereas Cluster 2 has mean values slightly below the full-sample averages. These profiles suggest an exploratory distinction between relatively higher-activity, higher-volatility and lower-activity, lower-volatility stock behavior. However, the overlap between the groups prevents their interpretation as sharply separated categories.

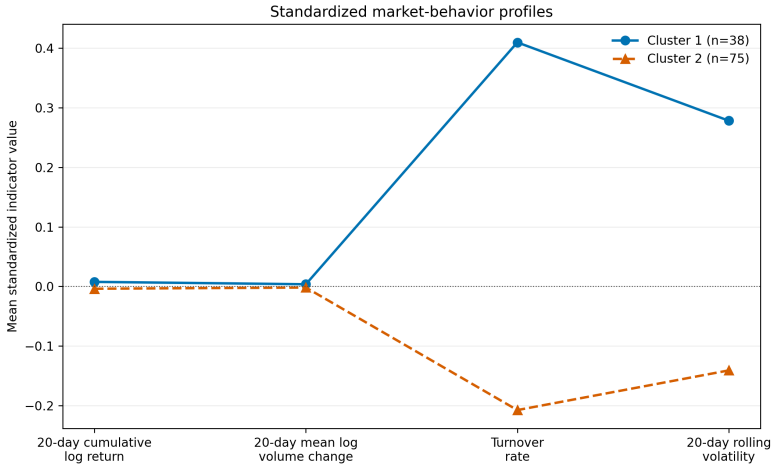


Figure 3. Standardized market-behavior characteristics of the stock clusters.

5.4. Comparative experiments and robustness analysis

5.4.1. Comparison with alternative representations

The two-cluster solution is selected relative to the candidate values $\kappa = 2, \ldots, 8$, not because its absolute separation is strong. To contextualize the modest silhouette coefficient, we compare K-means on the original standardized four-indicator series with PCA-K-means and the proposed spline-based representation. All four configurations use $\kappa = 2$ and 30 K-means restarts. PCA is fitted across the 113 stocks; the two PCA settings retain 100 components and 83 components (95.07% of explained variance), respectively. The proposed method uses 100 B-spline control points per indicator. Table 8 summarizes the resulting internal indices and the ARI relative to the proposed partition.

Table 8. Internal clustering indices and partition agreement for alternative representations.

Representation	Dimension	Silhouette	CH	DB	ARI vs. proposed
Raw standardized series	6028	0.1497	11.1700	3.0634	1.0000
PCA (100 components)	100	0.1473	11.3360	3.0512	0.9293
PCA (95% variance; 83 components)	83	0.1658	11.8030	2.9761	1.0000
LSPIA–B-spline (proposed)	400	0.1687	14.1668	2.7158	1.0000

The silhouette coefficients of all four representations are modest (0.1473–0.1687). Although the proposed representation has the highest silhouette and CH indices and the lowest DB index in this comparison, these measures do not establish strong separation or external validity. Moreover, PCA with 95% explained variance and the raw-series approach recover exactly the same partition as the proposed method (ARI = 1), whereas the 100-component PCA setting gives ARI = 0.9293. The ARI here measures agreement with the proposed partition, *not* accuracy against independently established behavioral labels. CH and DB are evaluated in different feature spaces and thus warrant cautious cross-representation interpretation. Collectively, these results indicate that the weak separation is not unique to the spline-based representation, but they do not identify its cause or prove that it is intrinsic to financial data.

5.4.2. Sensitivity to the number of control points

With the number of clusters fixed at $\kappa = 2$, as determined in Section 5.3, the sensitivity of the clustering results to the number of control points is further examined. The complete fitting and clustering procedure is repeated for $M = 80, 90, 100, 110, 120$. Taking the partition obtained at $M = 100$ as the reference, ARI is calculated to assess the agreement between the resulting partitions. The results are reported in Table 9.

Table 9. Sensitivity analysis with respect to the number of control points.

M	Feature dimension	Silhouette	CH	DB	ARI with baseline	Minimum cluster	Maximum cluster
80	320	0.2253	15.1495	2.5252	0.6663	28	85
90	360	0.2047	14.3566	2.6392	0.7608	31	82
100	400	0.1687	14.1668	2.7158	1.0000	38	75
110	440	0.1670	13.8469	2.7485	1.0000	38	75
120	480	0.1657	13.6503	2.7677	1.0000	38	75

At $M = 80$ and $M = 90$, the silhouette and CH indices are higher and the DB indices are lower than those obtained at $M = 100$. However, the corresponding ARI values of 0.6663 and 0.7608 indicate changes in stock assignments. When M increases to 110 or 120, the resulting partitions are identical to the baseline partition, with ARI values of 1.0000 and unchanged cluster sizes of 38 and 75.

The identical partitions at $M = 100, 110, 120$ show that $M = 100$ is the smallest tested control-point number within this stable-label range. Together with the fitting accuracy and compression results in Section 5.2, this finding further supports the selection of $M = 100$ as the baseline setting. The observed label stability is specific to the tested range of control-point numbers and does not imply that the stock partition is invariant under arbitrary changes in M.

5.4.3. Indicator ablation analysis

To assess the practical contribution of each of the four indicators to the clustering result, one indicator is removed at a time, after which full-sample standardization, LSPIA fitting, equivalent-feature construction, and K-means clustering are repeated. The resulting partitions are compared with that of the complete four-indicator model using ARI. Let $\boldsymbol{\ell}^{\text{full}}$ and $\boldsymbol{\ell}^{(-j)}$ denote the clustering labels obtained from the complete model and the model without indicator j, respectively. The corresponding agreement is measured by $\text{ARI}(\boldsymbol{\ell}^{\text{full}}, \boldsymbol{\ell}^{(-j)})$. The results are reported in Table 10.

Table 10. One-indicator-at-a-time ablation results.

Indicator combination	Silhouette	CH	DB	ARI	Minimum cluster size
Complete four-indicator model	0.1687	14.1668	2.7158	1.0000	38
Without 20-day cumulative log return	0.2218	17.0121	2.4590	0.7616	33
Without 20-day average volume-change rate	0.2378	16.0483	2.4622	0.6663	28
Without turnover rate	0.1715	10.2047	2.8166	0.4155	23
Without 20-day rolling volatility	0.1684	15.1986	2.6226	0.8950	41

It is evident from Table 10 that removing turnover rate produces the lowest ARI (0.4155), indicating the largest change in cluster assignments among the four ablation cases. The CH index also decreases from 14.1668 to 10.2047, while the DB index increases from 2.7158 to 2.8166. These results suggest that turnover rate has a substantial influence on the present clustering structure. Removing the volume-change indicator increases the silhouette coefficient to 0.2378 and yields an ARI of 0.6663. This indicates that the reduced representation has more favorable internal separation under the reported metrics, but also produces a different stock partition.

Similarly, removing cumulative log return improves the internal indices while yielding an ARI of 0.7616, indicating that return information affects the assignments of some stocks. Removing rolling volatility produces the highest ARI among the ablation cases (0.8950), suggesting a relatively limited effect on the cluster assignments. Nevertheless, rolling volatility remains relevant for describing the risk characteristics of the groups.

Overall, the complete four-indicator model does not outperform every reduced model in all internal clustering indices. The ablation results instead reveal heterogeneous indicator contributions: turnover rate has a pronounced effect on stock assignments, while the other indicators provide complementary information on returns, volume changes, and risk. Because ablation changes the feature space and its dimensionality, the internal indices should be interpreted as diagnostics of the resulting partitions rather than as direct causal measures of indicator importance.

5.4.4. Sensitivity to indicator construction and standardization

The baseline uses a 20-trading-day horizon for cumulative log return, average logarithmic volume change, and rolling volatility. To examine the effect of this choice, we additionally consider $h \in \{10, 20, 40\}$. For each horizon, these three indicators are reconstructed from the original daily records, while the contemporaneous turnover-rate definition remains unchanged. All three settings use the intersection of retained stocks and valid trading dates so that the comparison is not confounded by different samples. The four indicators are standardized anew within each setting, and the same cubic B-spline degree, $M = 100$ control points per indicator, $\kappa = 2$ clusters, and K-means configuration are

used. We report the silhouette, CH, and DB indices, cluster sizes, and the adjusted Rand index relative to the 20-day partition estimated on this common sample. This experiment varies the financial-indicator construction horizon, not the length of the overall stock-price sample. In the present experiment, the 20-day reference is nevertheless recomputed within the controlled comparison rather than copied from the original experiment.

All three settings retain the same 113 stocks and 1507 common trading dates (November 12, 2018–December 26, 2025). As shown in Table 11, the 10-day and 40-day windows yield higher silhouette coefficients (0.2124 and 0.2136, respectively) than the 20-day window (0.1643). Their CH and DB indices also indicate more compact or separated groups under the respective feature representations. However, the ARI values of 0.7617 and 0.6678 show that changing the indicator horizon alters some stock assignments. The 20-day setting is retained as a conventional approximate one-month horizon for the baseline analysis, not because it optimizes the internal clustering indices. Accordingly, the two-cluster partition should be treated as an exploratory, window-dependent behavioral grouping. The 20-day result in this controlled re-estimation (silhouette = 0.1643, cluster sizes 39/74) is reported separately from the original baseline (silhouette = 0.1687, cluster sizes 38/75); these values should not be interchanged before the difference between the two processing runs is reconciled.

Table 11. Sensitivity of stock-behavior clustering to the financial-indicator window length.

Window h	Silhouette ↑	CH ↑	DB ↓	Cluster sizes	ARI vs. $h = 20$
10	0.2124	15.4847	2.5619	32/81	0.7617
20	0.1643	14.1379	2.7232	39/74	1.0000
40	0.2136	14.9497	2.5470	29/84	0.6678

To assess the influence of preprocessing on stock clustering, three standardization schemes are compared: full-sample, stock-specific, and time-specific (cross-sectional) standardization. All experiments use the same 113 stocks, 1507 common trading days, four financial indicators, $M = 100$ control points per indicator, and $\kappa = 2$ clusters. The subsequent LSPIA fitting, feature transformation, and K-means clustering procedures remain unchanged.

The ARI is calculated relative to the full-sample standardization result to quantify agreement between the resulting partitions. The results are reported in Table 12.

Table 12. Robustness of stock-behavior clustering to standardization schemes.

Standardization	Silhouette	CH	DB	Cluster sizes	ARI
Full-sample (baseline)	0.1687	14.1668	2.7158	38/75	1.0000
Stock-specific	0.0713	9.7959	3.3444	53/60	0.2482
Time-specific (cross-sectional)	0.1650	11.3545	2.9568	32/81	0.7297

Under full-sample standardization, the silhouette coefficient and CH index are 0.1687 and 14.1668, respectively, while the DB index is 2.7158. Stock-specific standardization yields a lower silhouette coefficient of 0.0713 and an ARI of 0.2482, indicating substantial changes in the stock partition after removing stock-specific location and scale differences. Time-specific standardization produces a silhouette coefficient of 0.1650 and an ARI of 0.7297, indicating partial agreement with the baseline partition while retaining noticeable differences in individual stock assignments.

These findings support the use of full-sample standardization for the present objective of retaining both temporal variation and persistent cross-sectional differences. However, they also demonstrate that

the resulting stock groups depend on the preprocessing scheme. Since internal indices are calculated in different standardized feature spaces, their numerical values should not be interpreted as an absolute comparison of clustering quality across all preprocessing choices.

5.5. Discussion and limitations

The existing sensitivity analyses provide a further qualification. Under the controlled 10/20/40-day indicator-window experiment, silhouette values are 0.2124/0.1643/0.2136 and the alternative partitions have ARI values of 0.7617 and 0.6678 relative to its re-estimated 20-day solution (Table 11). The 20-day result in that experiment differs slightly from the original baseline and is not used interchangeably with it. Likewise, changing the standardization scheme changes cluster assignments (Table 12). These findings indicate sensitivity to representation choices, rather than stability of a universal two-group structure. Since all 113 stocks belong to the pharmaceutical sector, their common sector membership cannot serve as a two-class ground truth for behavioral clustering. Independently defined behavioral labels and alternative distance/clustering models remain useful directions for external validation. Accordingly, the two clusters are reported as a weakly separated yet interpretable exploratory partition, not as definitive market categories.

The empirical evidence is restricted to 113 pharmaceutical-sector A-share stocks observed on 1507 common trading days from November 12, 2018, to December 26, 2025. Accordingly, the reported two-cluster solution describes within-sector behavior in this particular market and sample period; it is not evidence of a universal partition across sectors or markets. Sector-specific differences in turnover, volatility, liquidity, and trading conventions may affect both the indicator distributions and the resulting distances. Application to another sector or market would require reconstructing the four indicators, applying an appropriate common basis and standardization protocol, and re-evaluating the cluster number and validity indices rather than transferring the present labels directly.

The 10-, 20-, and 40-day sensitivity experiment in Table 11 tests the indicator construction horizon on the same 113 stocks and 1507 aligned dates. It does not test shorter or longer historical sample periods, temporal transferability, or out-of-sample performance. A future temporal validation should estimate scaling parameters, fit the spline representation, and select clustering settings using a training period only; subsequent non-overlapping periods could then be represented with the training-period basis and assigned to the learned centroids. The fraction of reassigned stocks, agreement with independently re-estimated partitions, and stability of behavioral profiles would provide complementary evidence. Comparable analyses on other sectors and markets are also needed. No such out-of-sample or cross-market results are claimed in the present study.

In summary, the LSPIA-B-spline representation compresses each 1507×4 standardized series into a 100×4 control-coefficient matrix at the baseline setting, while the resulting two-group partition remains exploratory. The comparisons and sensitivity experiments show that the observed grouping depends on representation and preprocessing choices; its economic interpretation is therefore restricted to the present sample and experimental settings.

6. Conclusions

This paper proposes a stock-behavior clustering method based on an LSPIA-B-spline control-coefficient representation. Periodic return, risk exposure, and trading activity are used as the principal

descriptive dimensions. Four indicators, the 20-day cumulative log return, the 20-day average rate of change in trading volume, the turnover rate, and the 20-day rolling volatility, are constructed, and full-sample standardization is applied to eliminate differences in units and numerical scales.

For time-series representation, LSPIA is used to fit the multi-indicator time series of each stock with B-splines by least squares, thereby compressing the original high-dimensional time series. Because all stocks use the same B-spline basis, their control coefficients share a common functional meaning and are cross-sectionally comparable, providing a consistent basis for distance calculation and clustering.

For stock-similarity measurement, a functional distance is defined from the fitted B-spline series generated by the control coefficients. The B-spline collocation matrix is then used to express this functional distance equivalently as the Euclidean distance between transformed control-coefficient features. This equivalence preserves the function-space structure induced by the local support and overlap of B-spline basis functions, while allowing the coefficient features to be used directly by conventional clustering methods. It therefore provides K-means with a Euclidean feature representation that retains the structure of the B-spline function space.

For clustering analysis, the transformed control-coefficient features are used as the input to K-means, and K-means++ is adopted to select the initial cluster centers. To reduce the influence of random initialization, each candidate cluster number is evaluated through 30 independent runs, and the run with the minimum within-cluster sum of squares is retained. By jointly considering the silhouette coefficient, CH index, DB index, and cluster sizes, the final number of clusters is determined to be two.

Overall, this paper establishes a complete analytical framework encompassing financial-indicator construction, feature standardization, LSPIA-B-spline fitting, compressed control-coefficient representation, functional-distance calculation, stock clustering, and result evaluation. Based on dynamic characteristics in periodic return, risk fluctuation, and trading activity, the method identifies groups of stocks with similar market-behavior patterns within a sector. The results provide a new functional-data-analysis perspective for stock-market structure analysis, portfolio construction, risk identification, and sector-rotation research. Future work may further validate stock-market behavior structures by incorporating temporal subwindow experiments, firm fundamental information, and a more refined indicator system.

Author contributions

Siyu Shen: Conceptualization, Writing–original draft; Lijuan Hu: Software, Methodology, Writing–review & editing; Chengzhi Liu: Supervision, Writing–review & editing. All authors have read and approved the final manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the Natural Science Foundation of Hunan Province under grant number 2025JJ70357 and the College Student Innovation and Entrepreneurship Training Program Project, Hunan College of Humanities, Science and Technology.

Data availability

The data that support the findings of this study are available from public financial data platforms. Detailed data-processing procedures and code will be made available upon reasonable request.

References

1. R. N. Mantegna, Hierarchical structure in financial markets, *Eur. Phys. J. B*, **11** (1999), 193–197. https://doi.org/10.1007/s100510050929
2. L. J. Lao, Y. M. Shao, Dynamic clustering analysis of return series of industrial indexes in Chinese stock market (in Chinese), *J. Finance Econ.*, **30** (2004), 76–83.
3. T. W. Liao, Clustering of time series data–a survey, *Pattern Recognit.*, **38** (2005), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025
4. S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering–a decade review, *Inf. Syst.*, **53** (2015), 16–38. https://doi.org/10.1016/j.is.2015.04.007
5. R. Cerqueti, P. D'Urso, L. De Giovanni, M. Giacalone, R. Mattera, Weighted score-driven fuzzy clustering of time series with a financial application, *Expert Syst. Appl.*, **198** (2022), 116752. https://doi.org/10.1016/j.eswa.2022.116752
6. J. V. Sáenz, F. M. Quiroga, A. F. Bariviera, Data vs. information: using clustering techniques to enhance stock returns forecasting, *Int. Rev. Financ. Anal.*, **88** (2023), 102657. https://doi.org/10.1016/j.irfa.2023.102657
7. M. Kahssay, S. Miah, A comparative analysis of stochastic models for stock price forecasting: the influence of historical data duration and volatility regimes, *Quant. Finance Econ.*, **9** (2025), 602–630. https://doi.org/10.3934/QFE.2025021
8. A. López-Oriona, J. A. Vilar, Quantile cross-spectral density: a novel and effective tool for clustering multivariate time series, *Expert Syst. Appl.*, **185** (2021), 115677. https://doi.org/10.1016/j.eswa.2021.115677
9. R. Mattera, G. Scepi, P. Kaur, Time series clustering for high-dimensional portfolio selection: a comparative study, *Soft Comput.*, **29** (2025), 4219–4231. https://doi.org/10.1007/s00500-025-10656-2
10. P. D'Urso, L. A. García-Escudero, L. De Giovanni, V. Vitale, A. Mayo-Iscar, Robust fuzzy clustering of time series based on B-splines, *Int. J. Approx. Reason.*, **136** (2021), 223–246. https://doi.org/10.1016/j.ijar.2021.06.010

11. G. H. Tang, L. Zhu, C. F. Liao, F. W. Jiang, Asset pricing research based on autoencoder machine learning: a financial big data analysis perspective of the Chinese stock market (in Chinese), *J. Manag. Sci. China*, **27** (2024), 82–97. https://doi.org/10.19920/j.cnki.jmsc.2024.09.006
12. Q. M. Luan, J. Hamp, Automated regime classification in multidimensional time series data using sliced Wasserstein k-means clustering, *Data Sci. Finance Econ.*, **5** (2025), 387–418. https://doi.org/10.3934/DSFE.2025016
13. C. De Boor, *A practical guide to splines*, New York: Springer, 1978.
14. C. Y. Deng, H. W. Lin, Progressive and iterative approximation for least squares B-spline curve and surface fitting, *Comput. Aided Des.*, **47** (2014), 32–44. https://doi.org/10.1016/j.cad.2013.08.012
15. H. W. Lin, T. Maekawa, C. Y. Deng, Survey on geometric iterative methods and their applications, *Comput. Aided Des.*, **95** (2018), 40–51. https://doi.org/10.1016/j.cad.2017.10.002
16. D. Rios, B. Jüttler, LSPIA, (stochastic) gradient descent, and parameter correction, *J. Comput. Appl. Math.*, **406** (2022), 113921. https://doi.org/10.1016/j.cam.2021.113921
17. S. Sajavičius, Hyperpower least squares progressive iterative approximation, *J. Comput. Appl. Math.*, **422** (2023), 114888. https://doi.org/10.1016/j.cam.2022.114888
18. N. C. Wu, C. Z. Liu, Asynchronous progressive iterative approximation method for least squares fitting, *Comput. Aided Geom. Des.*, **111** (2024), 102295. https://doi.org/10.1016/j.cagd.2024.102295
19. N. C. Wu, C. Z. Liu, Randomized progressive iterative approximation for B-spline curve and surface fittings, *Appl. Math. Comput.*, **473** (2024), 128669. https://doi.org/10.1016/j.amc.2024.128669
20. C. Z. Liu, N. C. Wu, J. C. Li, Efficient two-dimensional randomized progressive iterative approximation for large-scale B-spline fitting, *CSIAM Trans. Appl. Math.*, **7** (2026), 556–575. https://doi.org/10.4208/csiam-am.SO-2025-0062
21. N. C. Wu, H. H. Cao, C. Z. Liu, Momentum-accelerated randomized geometric iterative methods for curve and surface approximation, *Comput. Aided Des.*, **192** (2026), 104011. https://doi.org/10.1016/j.cad.2025.104011
22. J. M. Karpoff, The relation between price changes and trading volume: a survey, *J. Financ. Quant. Anal.*, **22** (1987), 109–126. https://doi.org/10.2307/2330874
23. H. W. Xing, W. Q. Liu, Turnover: liquidity or uncertainty, *J. Shanghai Univ. Finance Econ.*, **20** (2018), 58–71. https://doi.org/10.16538/j.cnki.jsufe.2018.05.005
24. T. G. Andersen, T. Bollerslev, F. X. Diebold, P. Labys, Modeling and forecasting realized volatility, *Econometrica*, **71** (2003), 579–625. https://doi.org/10.1111/1468-0262.00418
25. S. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inf. Theory*, **28** (1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489
26. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, **20** (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

27. T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Commun. Stat.*, **3** (1974), 1–27. https://doi.org/10.1080/03610927408827101

28. D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1** (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

29. L. Hubert, P. Arabie, Comparing partitions, *J. Classif.*, **2** (1985), 193–218. https://doi.org/10.1007/BF01908075



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)