Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 4830716, 13 pages
https://doi.org/10.1155/2023/4830716

Research Article

Novel Approaches to Identify Clusters Using Independent Components Analysis with Application

Saima Afzal,[1] **Muhammad Mutahir Iqbal**,[1] **Ayesha Afzal**,[2] **Hassan S. Bakouch**,[3,4] **and Sadiah M. A. Aljeddani**[5]

[1]*Department of Statistics, Bahauddin Zakariya University, Multan 60000, Pakistan*
[2]*Department of Computer Science, Air University Multan Campus, Multan 60000, Pakistan*
[3]*Department of Mathematics, College of Science, Qassim University, Buraydah, Saudi Arabia*
[4]*Department of Mathematics, Faculty of Science, Tanta University, Tanta, Egypt*
[5]*Department of Mathematics, Al-Lith College, Umm Al-Qura University, Al-Lith, Saudi Arabia*

Correspondence should be addressed to Hassan S. Bakouch; hassan.bakouch@science.tanta.edu.eg

Received 7 July 2022; Revised 11 October 2022; Accepted 14 October 2022; Published 31 January 2023

Academic Editor: Tahir Mehmood

As a statistical and computational technique, independent component analysis (ICA) is employed to separate the source variables into statistically independent components. ICA methods have received growing attention as effective data mining tools. In this paper, two novel ICA-based approaches are proposed to identify the clusters of variables. The identified clusters reduce the dimensionality of the data in a natural way. The first approach, namely "Estimated Mixing Coefficients," is based on the sum of squares of mixing coefficients, and the second approach, namely "Ranked $\overline{R}^2$," uses the ranking pattern of $\underline{R}^2$ of the original and reconstructed series at predefined threshold levels. The proposed techniques are applied to financial time series data to validate their effectiveness. The main focus of the study is on the clustering of multivariate time series datasets using two new proposed approaches based on independent component analysis. The internal and external structures of clusters are also explored using different metrics. Both proposed techniques are compared with some existing clustering techniques. The experimental evaluation results show that the performance of the proposed techniques is better than the existing techniques.

1. Introduction

Clustering, as a dimension reduction technique, is quite helpful in deciding the number and structure of the classes. These classes are suitable representatives of the data, which are internally maximally homogeneous. The mutually exclusive and collectively exhaustive classes are termed clusters. Clustering is particularly useful in exploratory data analysis for summarization and as a preprocessing step in complex data mining tasks. In general, an effective clustering scheme produces internally homogeneous but externally heterogeneous clusters of sufficiently large size without using any prior knowledge of data divisions. The produced clusters, therefore, may differ from any theoretical division already available for the data [1].

A variety of clustering algorithms has been proposed in the literature. Clustering algorithms are typically categorized into partitioning methods, hierarchical methods, density-based methods, and grid-based methods [2, 3]. Most of the algorithms are developed for clustering of observations rather than dimensions or variables in a multivariate dataset. Our focus in this research is on the clustering of multivariate time series datasets, and we present two new approaches for such a clustering based on independent component analysis.

The independent component analysis (ICA) has been used for clustering different kinds of data, e.g., in works by Keck et al. [4], Jamal and Kent [5], and Islam et al. [6]. The ICA is a statistical and computational technique in which the objective is to find a linear projection of the data in which the source signals or components are statistically independent or as independent as possible. Essentially, the ICA linearly transforms the data in a way that the resulting components can be grouped into clusters. Each component is dependent within a cluster and independent across clusters. Among its numerous applications, ICA is the most natural tool for blind source separation in instantaneous linear mixtures when the source signals are assumed to be independent. The main reason for the increased interest of researchers in ICA is mainly due to the plausibility of the statistical independence assumption in a wide variety of fields, including sales, finance, telecom, weather forecasting, and biomedical engineering.

In this work, we propose two ICA-based approaches for variable clustering. ICA supports cluster identification by reducing the dimensionality of the data in a natural way. The first approach, the "Estimated Mixing Coefficients Approach," is based on the sum of squares of mixing coefficients. The second approach, namely "Ranked $\underline{R}^2$" uses the ranking pattern of $\underline{R}^2$ of the original and reconstructed series at predefined threshold levels. In order to validate the performance of our proposed techniques, we applied these approaches to a financial time series dataset with the objective of exploring the internal and external structures of identified clusters.

Financial time series represent data on asset valuation as a function of time and usually include parameters, such as stock market index values, currency exchange rates, electricity prices, and interest rates. Data mining of financial time series has established very effective and useful results. Financial time series is affected by some underlying factors, such as news (good or bad), government interference, natural or artificial disasters, and political upheaval. These underlying factors affect the volatility of time series. Clustering could be very helpful in analyzing the time series of a group including several stocks. The analysis of the financial time series of a portfolio including several stocks, can be carried out by clustering the stocks. The performance of an investment portfolio is not necessarily determined by the stock that formulates the largest monetary share of the investment. ICA can be applied to discover the underlying or hidden components, factors (e.g., some good or bad news, government interference, any natural or man-made disasters, political disorder, and response to massive trading), and to remove any noise.

The performance of the proposed approaches is compared with two existing approaches, namely Ward's method and the average linkage method. The supremacy of the proposed approaches is confirmed by the findings of comparative evaluation.

The primary contributions of this paper are as follows:

(i) Development of two new approaches for clustering based on ICA, namely the estimated mixing coefficients-based approach and the ranked $\underline{R}^2$-based approach

(ii) Experimental validation of the effectiveness of proposed approaches by application on a financial time series dataset for clustering of stock returns

(iii) Finding interpretable factors for stock returns in terms of ICs

Now, we discuss the notation used in the paper and formally describe the basic ICA model.

Consider a multivariate time series, $x_{\mathrm{it}} = \{x_{1t}, x_{2t}, \ldots, x_{rt}\}$ with r random variables at some time point t, modeled as linear combinations of m random variables $s_{1t}, s_{2t}, \ldots, s_{mt}$ given by the following:

$$\begin{aligned} x_{\mathrm{it}} &= w_{i1}s_{1t} + w_{i2}s_{2t} + \ldots + w_{im}s_{mt}, \\ \forall i &= 1, 2, \ldots, r, m \le r. \end{aligned} \quad (1)$$

With each w_{ij}: $i = 1, \ldots, r$ and $j = 1, \ldots, m$ being some real unknown parameter.

By definition each s_{it} are statistically mutually independent and nonGaussian distributed components.

Using vector-matrix notation equation (1) can simply be written as follows:

$$X = WS, \quad (2)$$

where X is an $r \times t$ matrix of observations, W is an $r \times m$ matrix of unknown parameters and is called the "Mixing Matrix," and $\mathbf{S}$ is an $m \times t$ matrix of nonGaussian and mutually independent hidden components called independent components (ICs).

The main objective of ICA is to estimate from the given sample of observations X, the mixing matrix W as well as the independent components, S. Thus, ICA attempts to find a linear transformation of the data as follows:

$$\widehat{S} = AX, \quad (3)$$

where a demixing matrix A of size $r \times m$ is to be identified such that the components (rows) of $\widehat{S}$ become as independent of each other as possible. Principal components analysis (PCA) has been a very common practice for identifying clusters in multivariate data over the past more than two decades. There is also some work on clustering using ICA or hybrid approaches where ICs are computed after applying PCA. For example, Reza et al. [7] proposed an approach to identify clusters through PCs, ICs, and ICs after PCs. Islam et al. [6] compared clusters formed by ICs, PCs, and ICs after PCs using four simulated datasets and three real-life datasets.

Bach and Jordan [8] proposed an approach where a transformation was searched to fit the estimated sources to a forest-structured graphical model. The optimal transformation for the nonGaussian temporally independent case was obtained by a mutual information-based contrast function. That mutual information-based contrast function extends the contrast function used for the classical ICA.

Keck et al. [9] proposed an algorithm to cluster signals using the incomplete ICA. In this approach, first, the ICA is applied to the dataset without reducing the dimensions; then, in the second step, dimension reduction is performed for clustering using similarity in elements of the mixing matrix.

Keck et al. [10] employed the ICA to identify clusters from functional magnetic resonance imaging (fMRI) data. The idea is to identify clusters by comparing the ICs computed at different levels of reduced dimensions. First, a set of ICs is computed without reducing the dimensions of the data. In the next iteration, the second set of ICs is computed from the dataset with reduced dimensions. The approach employs PCA for dimension reduction. After comparing the results of each iteration, matching ICs are retained to form clusters.

In another work on multivariate time series clustering, Wu and Yu [11] first employ the FastICA algorithm to transform multivariate time series into ICs and then select the dominant ICs (based upon loadings). Clusters are then identified based on the similarity of the dominant ICs. We also use FastICA in addition to other algorithms for estimating the ICs and computing the mixing matrix. However, our approach can use any efficient ICA algorithm.

Based on the fact that departure from Gaussianity helps in calculating ICs, some attempts have also been made to reduce Gaussianity as much as possible. This departure from Gaussianity is common in real-life situations. Inducing nonGaussianity can maximize the absolute kurtosis, which leads to some approaches that move in the positive or negative direction of kurtosis to attain sub-Gaussianity or super-Gaussianity. If the distribution is super-Gaussian, then it is least likely to have more than one mode located, whereas sub-Gaussianity increases the chances of having more than one mode identified. Jamal and Kent [5] proposed a clustering technique based on the fact that the clusters are formed when kurtosis is usually negative, i.e., the distribution is sub-Gaussian. Using the sub-ICA algorithm, ICs can be obtained by minimizing kurtosis and increasing the chances of locating modes. The one-dimensional projection of the so-calculated ICs would suggest modes, and each mode will center a cluster. This is how clusters are formed in this approach.

Lu and Chang [12] proposed a hybrid sales forecasting scheme by combining the ICA, *K*-means clustering, and support vector regression (SVR). The proposed scheme first applies ICA to extract hidden information from the observed sales data. In the next step, the *K*-means clustering algorithm is applied to extracted features. The SVR forecasting models are applied as the last step to each group to generate final forecasting results. The proposed approach provides forecasting models based on ICA and *k*-means clustering.

Azam and Bouguila [13] proposed a speaker classification method based on supervised hierarchical clustering. A bounded generalized Gaussian mixture model with the ICA is used for statistical learning with some modifications in the clustering framework. Using the training data, the ICA mixture model is learned, and posterior probability is used to divide the training data into clusters. The researchers proposed a supervised hierarchical clustering approach, which could be a complex procedure because supervised learning is more complex as compared to unsupervised learning.

Nascimento et al. [14] proposed an ICA-based clustering approach, namely ICAclust, to cluster gene expression data. It is a two-step clustering method that relies upon ICA and a hierarchical method for clustering at the same time. The performance of the ICA-based clustering was compared with *k*-means clustering. Overall their proposed method performed better than the *k*-means clustering method, but it was also observed that it performed better for the small number of temporal observations.

Gultepe & Makrehchi [15] used *K*-means, spectral clustering, graph regularized non-negative matrix factorization, and *K*-means with principal components analysis algorithms. They applied blind source separation (BSS) using the ICA were used for each clustering algorithm. They evaluated the performance of their proposed method using six benchmark datasets, which include five image datasets used in object, face, digit recognition tasks, and one text document dataset used in topic recognition. It was concluded that maximum clustering performance in four out of six datasets was achieved by applying ICA BSS after the initial matrix factorization step. The main drawback of this approach is the processing speed of the similarity graph and the matrix factorization due to the initial eigendecomposition.

Durieux and Wilderjans [16] worked on three-way fMRI data. They proposed a two-step procedure. In the first step, the ICA was applied to extract functional connectivity patterns from the data, and in the second step, a clustering algorithm was applied to identify the clusters of patients with similar functional connectivity patterns. The approach suffers from a model selection problem. While conducting the simulation study, the true number of clusters was assumed, and for reduction using the ICA or PCA, the true number of components for the original data was known. Furthermore, the number of components for each patient's fMRI data was assumed to be the same for every patient. The optimal number of cluster components in the empirical application for a dataset is not known a priori and has to be determined by the researcher. Incorrect specification of the true number of components may negatively affect the identification of the true cluster organization for a given dataset.

Shahina and Kumar [17] proposed a clustering approach based on similarity, which grouped the sensor node with similar data as a cluster for combining data. After that, an algorithm is proposed which combines the data making use of ICA, which is applied on cluster head sensor nodes. Data combining procedure was implemented on clusters having similarity of data. The study did not gain much as a very slight improvement of results is achieved in terms of aggregation ratio when compared with existing systems of self-organizing map (SOM) and PCA-based aggregation.

Boonyakitanont et al. [18] presented a work that performs subject group identification, latent source magnetoencephalography (MEG) estimation, and discriminatory source visualization. They applied hierarchical clustering on principal components (HCPCs) to identify cluster subject groups, which were based upon cognitive scores, and the ICA was implemented on MEG-evoked responses in such a way that not only higher-order statistics but also sample dependence within sources was considered. The proposed approach is specific to identifying the clusters for MEG data.

Most of the existing ICA-based clustering techniques available in the literature are based upon loadings or estimated mixing coefficients of dominant ICs alone and do not take remaining loadings into account. The main disadvantage of choosing only the dominant components is that the remaining components often include some important information that is lost. Our proposed techniques are also built over the ICA. In our first estimated mixing coefficients approach, we utilize the information provided by all the ICs. In the second ranked $\overline{R}^2$ approach, we reconstruct the original series using dominant ICs only. The evaluation results show that considering all ICs significantly improves the clustering results.

After providing some background definitions, a formal statement of the problem, and a brief review of the related literature, the rest of the paper is organized as follows. Section 2 presents two new approaches to cluster the stock data. The application of the proposed approaches is presented in Section 3. Section 4 provides the analysis of the identified clusters. Finally, Section 5 concludes the paper.

2. The Proposed Clustering Approaches

In this section, we discuss in detail the two ICA-based approaches we have developed for clustering. The first approach utilizes all of the mixing coefficients and the second one is based on the reconstruction of the series with dominant ICs. The computation of ICs is the first step for both proposed approaches.

2.1. Computation of ICs. Many approaches exist in the literature for estimating ICs and the mixing matrix, including maximization of nonGaussianity, information theoretic measures, maximum likelihood estimation method, and tensor-based methods. In this work, we make use of three prominent algorithms proposed for specific applications to financial data [19, 20] including JADE, SOBI, and FastICA, for a comparative assessment. We have briefly discussed these algorithms in the article.

2.1.1. Joint Approximation Diagonalization of Eigenmatrices Algorithm (JADE). JADE [21] was developed following the seminal work of Back & Weigend [19], who first proposed ICA for exploring the structure of stock returns. JADE is based on higher-order statistics. Higher-order statistics-based algorithms rely on the characteristics of the data distribution to perform the separation. This makes such algorithms robust to additive Gaussian noise. The rule working behind the algorithm is the solution to the problem of equal eigenvalues of the cumulant tensor. The main quality of JADE is its computational efficiency for blind estimation of directional vectors, which is based on joint diagonalization of fourth-order cumulant matrices.

2.1.2. Fixed-Point Algorithm (FastICA). FastICA [22, 23] is also a higher-order statistic-based algorithm. It makes use of kurtosis for the estimation of ICs. Data whitening is a preprocessing step for the algorithm. Mainly, FastICA works on the principle of the maximization of nonGaussianity to obtain independence. FastICA is known to be computationally very efficient with parallel implementations. However, the main drawback of FastICA is the loss of temporal information and higher memory requirements in the case of nonparallel implementations.

2.1.3. Second-Order Blind Identification Algorithm (SOBI). SOBI [24] is based on second-order statistics. SOBI is a three-step algorithm that makes use of time-frequency information for decomposition. In the first step, data whitening is performed; in the next step, lagged correlation matrices are computed; and in the third step, blind source separation is performed by approximate joint diagonalization of time-delayed covariance matrices.

Table 1 summarizes the main features of the above three algorithms.

2.2. Estimated Mixing Coefficients Approach: The First Approach. Our first approach utilizes all the mixing coefficients. Basically, this approach is based upon the reconstruction of variables with reduced dimensions and concentrates on the comparison of the ICs themselves. Algorithm 1 outlines the basic steps in our approach.

In the first step (line 1 of Algorithm 1), we compute the ICs for the given input series given as an $r \times t$ matrix X using any ICA algorithm, such as FastICA, JADE, or SOBI. Let, the matrix of ICs be given by the following:

$$S = \left[s_{\text{it}}\right] = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1,T} \\ s_{21} & s_{22} & \cdots & s_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,T} \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix}. \quad (4)$$

In the second step (lines 2 and 3 of Algorithm 1), we compute the estimated mixing matrix W and the corresponding separating matrix A given as follows:

$$W = \left[w_{ik}\right] = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,r} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,r} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}, \quad (5)$$

$$A = W^{-1} = \left[a_{ik}\right] = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,r} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,r} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}. \quad (6)$$

As discussed by Back & Weigend [19]; for $A = W^{-\mathbf{1}}$, we have three basic assumptions: (i) all sources s_{it} are statistically independent, (ii) at most one source has a Gaussian distribution, and (iii) the observations are stationary.

Note that when the ICA is applied for dimension reduction, one main issue is how to rank or order the ICs and rows of the mixing matrix in terms of significance to select dominant components. Our approach for obtaining such an ordering of ICs is to use the sum of squares of mixing

Table 1: Main features of the JADE, FastICA, and SOBI.

Category	Method	Algorithms	Pros	Cons
Higher order statistic-based approach	NonGaussianity	JADE	(i) Computationally efficient on the low dimensional datasets in terms of running time requirements	(i) Does not consider the temporal characteristics of the dataset
			(ii) Stable in terms of memory space requirements	(ii) Inefficient for high dimensional datasets in terms of computational speed.
		FastICA	(i) Computationally efficient in terms of running time	(i) Does not consider the temporal characteristics of the dataset
			(ii) Capability for parallel implementation	(ii) Not robust when criteria for nonGaussianity measurement is kurtosis (iii) Higher memory space requirements
Second-order statistic-based approach	Temporal dependence	SOBI	(i) Time-delayed covariance matrices of estimated independent components are closest to the diagonal (ii) Computationally efficient both in terms of memory space and execution time	Does not consider the selection of auto-covariance order

```
     Input: $X$: $r \times t$ matrix of observations, $k$: number of clusters
     Output: $C$: clusters
(1)  $S$ = computeICs($X$);/* execute ICA algorithm to generate ICs */
     /*compute estimated mixing matrix $W$ and the corresponding demixing matrix $A$*/
(2)  $W$ = computeMixingMatrix($X$);
(3)  $A$ = inverseMatrix($W$);
     /*compute sum of squares of mixing coefficients in $A$*/
(4)  for $i = 1$ to $m$
(5)      for $k = 1$ to $r$
(6)          sumVect[i] = sumVect[i] + squareOf($A[i][k]$)
(7)      end for
(8)  end for
(9)  sortAscending(sumVect);
     /* cluster the ordered rows in sumVect into $k$ clusters, using an arbitrary clustering scheme */
(10) $C$ = performClustering(sumVect, $k$)
```

Algorithm 1: EMC_clustering_algorithm.

coefficients and reordering the rows in the obtained sum of squares vector in ascending order.

Therefore, the next step in our approach is to compute the sum of squares of mixing coefficients in matrix A. For each row $\mathbf{a}_i$ of A, we compute the sum of squares, $\sum_{k=1}^{r} a_{ik}^2$ (lines 4–8 of Algorithm 1).

Finally, we partition the ordered rows obtained in the previous step into k equal sized clusters (line 8 of Algorithm 1). Several criteria are available in the literature to determine a reasonable k for clustering. We follow the criterion given by Mardia et al. [25]; i.e., $k \approx \sqrt{m/2}$, where k is the number of clusters and m is the number of objects/variables. Any robust criterion for determining the value of k may be adopted.

2.3. Ranked $\overline{R}^2$ Approach: The Second Approach. The key idea behind our second approach is to compare the reconstruction of the original variables at different threshold levels of dimension reduction.

The step-by-step procedural details of this approach are discussed as follows:

(1) Similar to our first approach, perform the ICA for the input series, given as an $r \times t$ matrix X to obtain the matrix of ICs as given by equation (4). Then compute the mixing and separating matrices, W and A, respectively.

(2) Arrange the computed ICs in an appropriate order. For this, we apply a regression-based method proposed by Afzal and Iqbal [26]. Given the m independent components and the mixing matrix, each row $i = 1, \ldots, m$ in the original series is regressed on all m independent components (here, we have $r = m$) to obtain all regression coefficients except the intercept.

   (i) Using the corresponding mixing matrix row for the i^{th} original series used above, rank 1 is assigned to an element of i^{th} row of the mixing matrix whose magnitude is closest to the magnitude of the first regression coefficient. The pair of the regression coefficient and the element of the i^{th} row of the mixing matrix, which has just been assigned rank 1 are set aside.

(ii) Similarly, rank 2 is assigned to the element of the i^{th} row of the mixing matrix whose magnitude is closest to the magnitude of the second regression coefficient. The second pair of regression coefficients and the element of the i^{th} row of the mixing matrix just ranked 2 are set aside.

(iii) The procedure is repeated till all the elements of the i^{th} row of the mixing matrix are ranked.

Assigned ranks are then used to arrange the corresponding ICs. The ICs matrix with ordered rows using the above process is given by

$$S' = s_{\text{it}}^{'} = \begin{bmatrix} s_{11}^{'} & s_{12}^{'} & \cdots & s_{1,t}^{'} \\ s_{21}^{'} & s_{22}^{'} & \cdots & s_{2,t}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m,1}^{'} & s_{m,2}^{'} & \cdots & s_{m,t}^{'} \end{bmatrix} = \begin{bmatrix} s_1^{'T} \\ s_2^{'T} \\ \vdots \\ s_t^{'T} \end{bmatrix}. \tag{7}$$

The mixing matrix with ordered rows is given by

$$A' = a_{ik}^{'} = \begin{bmatrix} a_{11}^{'} & a_{12}^{'} & \cdots & a_{1,m}^{'} \\ a_{21}^{'} & a_{22}^{'} & \cdots & a_{2,m}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}^{'} & a_{m,2}^{'} & \cdots & a_{m,m}^{'} \end{bmatrix} = \begin{bmatrix} a_{\ 1}^{'T} \\ a_{\ 2}^{'T} \\ \vdots \\ a_{\ m}^{'T} \end{bmatrix}. \tag{8}$$

(3) Reconstruct the series by using the Back and Weigend [19] procedure, and do the reconstruction of each of the series at different arbitrary threshold levels (say p). Weighted ICs and threshold ICs are computed to reconstruct the series.

The matrix W' of weighted ICs is computed by using the procedure followed by Back and Weigend [19]. The elements of the i^{th} row are used as weights to compute weighted ICs. For the i^{th} variable the weighted ICs are computed by multiplying $a_{i1}^{'}$ (which is a scalar quantity) to $\mathbf{s}_1^{'}$ vector, $a_{i2}^{'}$ to $\mathbf{s}_2^{'}$ and so on. The matrix of weighted ICs is given as follows:

$$W' = \left[a_{ik}^{'} s_{kt}^{'}\right] = \begin{bmatrix} a_{i1}^{'} s_{11}^{'} & a_{i1}^{'} s_{12}^{'} & \cdots & a_{i1}^{'} s_{1,t}^{'} \\ a_{i2}^{'} s_{21}^{'} & a_{i2}^{'} s_{22}^{'} & \cdots & a_{i2}^{'} s_{2,t}^{'} \\ \vdots & \vdots & \ddots & \vdots \\ a_{im}^{'} s_{m1}^{'} & a_{im}^{'} s_{m2}^{'} & \cdots & a_{i,m}^{'} s_{m,t}^{'} \end{bmatrix}. \tag{9}$$

The threshold ICs are also computed by following Back and Weigend [19]. An arbitrary threshold level is used here. The $(m-l)$ components from bottom are excluded where m is the total number of components and l is the number of components to be retained.

For the i^{th} variable, and at time t, the threshold IC is computed as $\widehat{x}_{\text{it}} = \sum_{k=1}^{l} a_{ik}^{'} s_{kt}^{'}$. (Here $\widehat{x}_{\text{it}}$ is an estimated value of x_{it}).

(4) Use the original series as original data points and reconstructed series as fitted points and their difference as an error. Note that we need to summarize how close the original and reconstructed series are.

(5) Compare each of the reconstructed series with the original series and compute the adjusted coefficient of determination $(\overline{R}^2)$

(6) For each of the given series, p values of $\overline{R}^2 s$ are available. Rank these p values of $\overline{R}^2$ in ascending order for each variable.

(7) Check the ranking patterns of all variables to find similarities. Form clusters of the variables with similar ranking patterns. This automatically defines the number as well as the internal structure of the cluster.

Algorithm 2 outlines our ranked $\overline{R}^2$ approach.

3. Application of the Proposed Approaches

We apply the ICA for analyzing financial time series data of the Karachi Stock Exchange 100 Index (KSE-100 index) in order to measure the effectiveness of the proposed methods for clustering variables. An effective time-series clustering can be achieved if and only if the price fluctuations of stocks within a group or cluster are maximally correlated, but the price fluctuations of stocks between different groups are uncorrelated [1]. This is the key assumption that forms the basis of clustering stocks data.

The KSE-100 index is a benchmark for comparing stock price performance in Pakistan over a period of time. The dataset covers the daily closing rates of 161 companies of KSE for the period of June 11, 2004, to February 15, 2012. Each of the 161 companies consists of 2004 observations . Rates for the closed market days (other than Saturday and Sunday) are taken on the basis of the last day's closing rates.

Let the matrix of closing rates of 161 companies at 2004 time points be given by

$$Y = y_{it} = \begin{bmatrix} y_{10} & y_{11} & \cdots & y_{1,2003} \\ y_{20} & y_{21} & \cdots & y_{2,2003} \\ \vdots & \vdots & \vdots & \vdots \\ y_{161,0} & y_{161,1} & \cdots & y_{161,2003} \end{bmatrix} = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_{161}^T \end{bmatrix}. \tag{10}$$

Each value of y_{it} and $y_{\text{it}-1}$ denote the closing rates of i^{th} company's stock for two sequential days in the market.

3.1. Preprocessing. Stationarity is a standard requirement for most modeling approaches including the ICA. Note that stationary signals have a constant expected value which is not the case with stock prices. Therefore, we first convert the nonstationary stock prices, i.e., the closing rates, y_{it} (where $i = 1, 2, \ldots 161$ and $t = 1, 2, \ldots, 2003$) to stock returns. This is typically accomplished by taking the difference between consecutive values of the stock prices as the change in stock prices is relatively higher over the years [19]. Therefore, we compute relative returns to obtain a transformed stationary

```
Input: $X$: $r \times t$ matrix of observations
Output: $C$: clusters
$S = \text{computeICs}(X)$;/* execute ICA algorithm to generate ICs */
/*compute estimated mixing matrix W and the corresponding de-mixing matrix $A$*/
$W = \text{computeMixingMatrix}(X)$;
$A = \text{inverseMatrix}(W)$;
/* determine a ranking of ICs in $S$*/
$S' = \text{orde rICs}(S)$;
Perform reconstruction of the original series at arbitrary threshold levels ($P = \{p_i: i = 1\text{tor}\}$)
Compare each of the reconstructed series with original series and compute adjusted coefficient of determination ($\overline{R}^2$)
Rank the $p$ values ($\in P$) of computed $\overline{R}^2 s$ in ascending order for each variable in $X$.
/ *Perform clustering of variables based on similar ranking patterns.*/
$C = \text{performClustering}(X, P)$;
```

Algorithm 2: RR2_clustering_algorithm.

series x_{it} (where $i = 1, 2, \ldots, 161$ and $t = 1, 2, \ldots, 2003$) by describing geometric growth taking instead of additive for the sake of efficiency using equation (11) as follows:

$$x_{\text{it}} = \ln\left(\frac{y_{\text{it}}}{y_{i,t-1}}\right). \tag{11}$$

The matrix of transformed series (relative returns) is given by

$$X = x_{it} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,2003} \\ x_{21} & x_{22} & \cdots & x_{2,2003} \\ \vdots & \vdots & \ddots & \vdots \\ x_{161,1} & x_{161,2} & \cdots & x_{161,2003} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{161}^T \end{bmatrix}. \tag{12}$$

3.2. Application of the Estimated Mixing Coefficients Approach. The ICA is applied to 161 mixed signals, i.e., stock returns of companies, in this case, each having a sample size of 2003. As discussed earlier, this approach is based upon the sum of squares of mixing coefficients. Different algorithms to compute ICs produce a different matrix of mixing coefficients, but every algorithm produces the same sum of squares of mixing coefficients of rows; therefore, any of the algorithms discussed in Section 4.1 can be used.

Following the main assumption in the experimental setting for financial time series dataset analysis by Back and Weigend [19], we also assume that the number of mixed signals is equal to the number of source signals in all the experiments. Here, we have 161 mixed signals (companies), each having 2003 stock returns.

Algorithm 1 is supplied with all 161 stocks as input. The ICA algorithm returns 161 source signals in the form of ICs. Matrix of estimated ICs, **S**, and the estimated mixing matrix, **A**, are obtained. The number of clusters is defined using the rule given by Mardia et al. [25]. In our case of the 161 stock companies' dataset, this rule suggests over 9 clusters which are rounded to 10.

The sum of squares $\sum_{k=1}^{161} a_{ik}^2$ for each row $\mathbf{a}_i$ of $\mathbf{A}$ is computed. The rows denoting different companies are reordered in ascending order of the sum of squares. The ordered rows (companies) are divided into equal parts. The 161 companies are then divided into nine groups, each of size 16. The tenth group is of size 17. Each group is considered a cluster. Application of the estimated mixing coefficients approach on this dataset returns the clusters presented in Table 2.

3.3. Application of Ranked $\overline{R}^2$ Approach. The rows of the matrix of ICs and mixing matrix, as given in equations (5) and (7), are ordered using the regression-based ordering method proposed by Afzal and Iqbal [26]. The original series are reconstructed with reduced dimensions following the Back and Weigend [19] procedure. The reconstruction of the original series is performed at nine different threshold levels, i.e., using 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent of ICs for the purpose following Afzal et al. [27].

Let $\hat{\mathbf{Y}}$ be the matrix of the reconstructed series of closing rates of 161 companies of KSE, each having 2004 observations, then for a given retention level using equation (11), we have the following relationship:

$$\widehat{Y} = \left\{\widehat{y}_{i(t-1)}\right\} \cdot \left\{\text{antilog}\left(\widehat{x}_{\text{it}}\right)\right\}. \tag{13}$$

Using this relationship, $\widehat{y}_{i0}$ is required to proceed any further, which is borrowed as the starting point from the original series.

Each of the reconstructed series is then compared with the original series and using the original series, as original data points and reconstructed series as fitted points and their difference as error, $\overline{R}^2$ is calculated. Thus, nine values of $\overline{R}^2$ for every company are obtained. For a given company, say ABOT, nine values of $\overline{R}^2 s$ are available. These nine values of $\overline{R}^2$ are ranked in ascending order for each company. The ranking pattern of all the companies is checked to find similarities. Clusters of companies sharing similar patterns are formed. This automatically defined the number and size of clusters. The identified clusters based upon ranked $\overline{R}^2$ approach are presented in Tables 3–5 for JADE, FastICA, and SOBI algorithms, respectively. Twenty-two clusters were identified using JADE and FastICA each, and 10 clusters using the SOBI algorithm.

4. Analysis of the Quality and Structure of Clusters

In this section, we analyze the internal structure of the clusters returned by the proposed approaches by first comparing them with the sectors defined by KSE, and then we check the validity of clusters, i.e., exploring them on their own.

4.1. Structural Comparison with Sectors Defined by KSE. The KSE has defined 33 sectors altogether based on the primary activities of listed companies. In this section, we compare the clusters returned by the proposed approaches with the sector-wise grouping provided by KSE. Table 6 shows the grouping of the 161 companies in our dataset in these sectors.

It can be viewed from Tables 2–4 that five sectors (mentioned in Table 5) including commercial banks, nonlife insurance, life insurance, financial services, and equity investment instruments (S21, S22, S23, S24, and S25), are moving together and form a group which we term as "Money & Bank" group.

4.1.1. Clustering by the Estimated Mixing Coefficients Approach. If individual clusters formed by the estimated mixing coefficients approach in Table 2 are analyzed, then it is apparent that six companies from the oil and gas sector (S1) and pharma and biotech Sector (S17) are combined in Cluster 1. Cluster 2 has a majority of companies from chemicals sector. Cluster 3 has six companies from money and bank Group. About half of the companies from the industrial metal and mining sector are part of cluster 4. Four companies from the sector general industries are gathered in cluster 6. Cluster 7 is comprised of 5 companies from money and bank group. The remaining clusters do not exhibit any such pattern.

4.1.2. Clustering by the Ranked $\overline{R}^2$ Approach. The ranked $\overline{R}^2$ approach using the JADE algorithm returned 22 clusters, as shown in Table 3. The first cluster consists of 34 companies, of which six belong to the money and bank group, whereas one set of five are from the construction and material sector, another set of five are from the personal goods and textile sector, and the remaining 18 companies form smaller groups from other sectors such as food producers, automobile, and parts and chemicals. The first cluster thus takes the shape of a contrast where inversely related groups of sectors get put together, which negate each other in the sense that the positive behavior of the money and bank group, for example, causes a negative impact on the construction and material sector, that is people try to deposit money in the banks rather than consuming it on construction and material. The rationale visible in this cluster does not persist in the remaining clusters, so the argument cannot be forwarded ahead. Negation to the argument is quite obvious in the subsequent clusters, wherein in the second cluster, four out of twelve companies belong to money and bank group; two companies are from the automobile and parts sector. Similarly, four out of eleven companies in the third cluster are from money and bank group, and two companies are from the automobile and parts sector. In the fourth, cluster four out of nine companies belong to money and bank group. Two of them are from the food producers sector.

The results in Table 4 show that the calculation based upon FastICA produced similar results where, in the first cluster of size 39, ten companies belong to the money and bank group, six to the construction and materials sector, and three to the personal goods (textile) sector. The second cluster of 10 companies does not show a good internal structure as it includes three companies from the food producers sector and two from the chemical sector, whereas the remaining do not form any group. The third cluster is relatively smaller in size, where three companies are from the automobile and parts sector and two from money and bank group. Cluster 4 includes seven companies out of which three are from money and bank group and two from personal goods (textile) sector.

Among the three algorithms, SOBI's results presented in Table 5 are the worst in the sense that too many clusters of relatively very small size are formed. The largest cluster identified contains only five companies. That is, the whole spirit of clustering is ruined.

The comparison shows that the already defined sectors cannot be used as clusters. The discrepancy can be justified on the ground that the closing rates of the company do not follow a pattern governed by sectors rather they play their role independently. The stock market is based on perspicacity, which has become even more important in modern times because of online trading. Due to this, a lot of inexperienced day-traders have moved towards stock market trading; for example, HINO is classified by KSE in the engineering sector because it earns the largest portion of its revenue from this sector. If most of the investors recognize HINO as part of the automobile and parts sector then its price fluctuation will follow the behavior of the automobile and parts group. Clustering could also be very helpful in analyzing the time series of a group including several stocks. The behavior of an investment group is not necessarily determined by the stock that makes up the largest monetary share of the investment. Clustering the stock data could identify which groups have the greatest influence on the portfolio. It is difficult to identify the cluster of stocks as their appropriate sectors because of the uncertain behavior of some stocks. Similar results were presented by Wittman [28]. Thus clustering should not be confused with pre-defined grouping whatsoever.

The next section concentrates on the exploration of the internal structure of clusters on their own.

4.2. Validity of Clusters. In this section, we present an evaluation of the quality of clusters identified by our proposed clustering techniques by comparing them with the quality of clusters identified using two of the most widely used clustering methods, including Ward's method [29] and the average linkage method [30].

Table 2: The ICA-based clusters using the estimated mixing coefficients approach.

Cluster no.	Cluster size	Companies
1	16	HUBC, AGTL, OGDC, GLAXO, FFC, KOHE, NESTLE, ABOT, SHEL, IBFL, SITC, PSO, PKGS, INDU, HINOON, and GHGL
2	16	FFBL, SIEM, PAKT, PTC, ENGRO, ICI, PSMC, AGIL, CPL, POL, ACPL, FEROZ, MTL, SEARL, JDWS, and LUCK
3	16	BAHL, MEBL, IDYM, NRL, KSBP, GTYR, PAKD, PRL, ATBA, MCB, SCM, CEPB, MUREB, NBP, FHAM, and SEL
4	16	PNSC, PICT, GADT, ATLH, HINO, GLPL, NML, SHFA, CSAP, PECO, DGKC, HABSM, BIFO, CLOV, FABL, and SAZEW
5	16	HICL, PCAL, MIRKS, OLPL, PAEL, CHCC, BWHL, ALNRS, ATRL, MARI, DYNO, INIL, FCCL, HSPI, EFUL, and BNWM
6	16	SING, ADOS, CRTM, CENI, BOP, NCL, KOHC, PIOC, HCAR, RICL, HAL, AGIC, BYCO, EFUG, FECTC, and MLCF
7	16	FHBM, KASBB, PIAA, GHNL, REWM, ANL, KESC, FCSC, DAWH, LOTPTA, SANSM, KTML, SEPCO, GASF, SHSML, and NIB
8	16	CHAS, JOPP, GWLC, FNBM, FDIBL, FUDLM, ADMM, JOVC, SGML, DSFL, ECOP, JPGL, NATF, DFML, PAKRI, and PSYL
9	16	NICL, KOHP, JSCL, IDRT, MACFL, SAIF, ESBL, PMI, EMCO, FFLM, DWSM, KOIL, FEM, BGL, DNCC, and MZSM
10	17	PNGRS, PTEC, CPMFI, FECM, QUICE, FPJM, SIBL, RAVT, GENP, MODAM, FRCL, MFTM, HADC, PAKMI, COTT, AICL, and MUKT

Table 3: The ICA-based clusters using ranked $\overline{R}^2$ approach (case: JADE).

Cluster no.	Cluster size	Companies
1	34	ACPL, ALNRS, BAHL, BNWM, BWHL, CHCC, DAWH, DGKC, DNCC, DSFL, EFUG, ESBL, FABL, GLAXO, GTYR, HINO, HINOON, HUBC, IDRT, JDWS, KESC, KOIL, LUCK, MARI, MFTM, MIRKS, MLCF, MODAM, MUKT, OGDC, PAEL, PIAA, SAIF, and SCM
2	12	ATBA, BGL, GADT, GLPL, JOPP, JSCL, KASBB, NESTLE, OLPL, PAKMI, PAKT, and PSO
3	11	AGIC, CPL, FPJM, IBFL, ICI, IDYM, PAKRI, PKGS, RICL, SAZEW, and SEPCO
4	9	ATLH, CLOV, EFUL, FDIBL, FUDLM, HICL, MUREB, NATF, and SING
5	6	AGIL, FCCL, GHNL, KOHE, PIOC, and SHFA
6	6	FHAM, KSBP, NCL, SHEL, SIEM, and SITC
7	5	GWLC, HSPI, PAKD, PCAL, and SIBL
8	4	ANL, FECTC, FNBM, and PRL
9	3	ABOT, BYCO, and MZSM
10	3	ADOS, ATRL, and ECOP
11	3	BIFO, MCB, and SANSM
12	3	CPMFI, FEROZ, and FHBM
13	3	ENGRO, FFC, and PNGRS
14	3	GASF, PNSC, and SEARL
15	2	ADMM, and CEPB
16	2	AICL and FFLM
17	2	CENI and FRCL
18	2	CHAS and CRTM
19	2	CSAP and DFML
20	2	FECM and PMI
21	2	FEM and NIB
22	2	PICT and RAVT

Table 4: The ICA-based clusters using ranked $\overline{R}^2$ approach (case: FastICA).

Cluster no.	Cluster size	Companies
1	39	AGIC, AGTL, ANL, BNWM, CHCC, COTT, DNCC, DWSM, EFUL, ESBL, FABL, FCCL, FCSC, FECTC, FFLM, FHBM, GHNL, HCAR, HICL, HINO, HINOON, INIL, JDWS, JSCL, KASBB, KOHC, LUCK, MARI, MTL, NATF, NBP, PAKRI, PIAA, PNGRS, POL, PRL, SAZEW, SEARL, and SIEM
2	10	ABOT, AICL, CLOV, DAWH, ICI, IDRT, MIRKS, NESTLE, OLPL, and PSMC
3	8	ATLH, ATRL, BWHL, DFML, GASF, MCB, MODAM, and PTEC
4	7	CPL, EFUG, FUDLM, JOVC, NCL, PCAL, and REWM
5	6	BAHL, KOHP, KTML, MACFL, SAIF, and SANSM
6	5	GWLC, FNBM, NICL, PAKD, and QUICE
7	4	CENI, KSBP, PICT, and SIBL
8	4	DYNO, ENGRO, IDYM, and SHFA
9	4	FDIBL, FEM, JOPP, and PAEL
10	3	DGKC, JPGL, and MUREB
11	3	FFBL, GLAXO, and HADC
12	3	HSPI, NIB, and RICL
13	2	ACPL and MEBL
14	2	AGIL and GLPL
15	2	BYCO and DSFL
16	2	CEPB and GTYR
17	2	CHAS and GHGL
18	2	FPJM and PMI
19	2	HAL and SEPCO
20	2	HUBC and PKGS
21	2	KESC and PSO
22	2	MLCF and SCM

Table 5: The ICA-based clusters using ranked $\overline{R}^2$ approach (case: SOBI).

Cluster no.	Cluster size	Companies
1	5	HUBC, PAKT, PSO, PSYL, and SITC
2	4	AGTL, FECTC, HSPI, and NRL
3	3	ATBA, DGKC, and ICI
4	2	AICL and POL
5	2	BOP and NML
6	2	BWHL and NATF
7	2	EMCO and SAIF
8	2	GLPL and NBP
9	2	MIRKS and SANSM
10	2	PECO and SAZEW

The quality of clustering can be gauged by measuring the internal homogeneity and external heterogeneity of the clusters. The clustering of a financial time series can be considered credible only when the stock prices within a cluster are maximally correlated, but different clusters are minimally correlated [31].

Various indices are available in the literature to determine the validity of identified clusters. Two of the popular and fundamental ones are given as follows:

(i) Calinski–Harabasz Index (CHI): Calinski and Harabasz [32] introduced this index to assess the quality of the clustering solution by analyzing the similarity of the objects within each cluster and the dissimilarity of different clusters. This index is also called the variance ratio criterion (VRC). The larger value of CHI indicates better data partition. The CH index for K number of clusters on a data set $X = [x_1, x_2, \ldots x_N]$ is given as

$$\mathrm{CH} = \frac{\left[\sum_{k=1}^{K} n_k \left\| v_k - v \right\| / K - 1\right]}{\left[\sum_{k=1}^{K} \left\| x_i - v_k \right\| / N - k\right]}, \tag{14}$$

where, n_k and v_k are the number of points and centroid of the kth cluster, respectively. v is the overall centroid and N is the total number of data points.

(ii) Davies–Bouldin Index (DBI): this index was introduced by Davies and Bouldin [33] and is based on the ratio of within-cluster-distance to between-cluster-distance. The lower value of the index indicates a better cluster structure. The DBI is calculated for K clusters as follows:

$$\begin{aligned}
\mathrm{DB} &= \frac{1}{K} \sum_{k=1}^{K} R_k, \text{where,} \\
R_k &= \max\left(R_{kl}\right); \quad k, l = 1, 2, \ldots, K, k \neq l, \\
R_{kl} &= \frac{s_k + s_l}{d_{kl}}, \text{where,} \\
d_{kl} &= d\left(v_k, v_l\right), s_i = \frac{1}{\left\| c_k \right\|} \sum_{x \varepsilon c_k} d\left(x, v_k\right).
\end{aligned} \tag{15}$$

$d(x, z)$ is the Euclidean distance between x and z, c_k is the k^{th} cluster, v_k is the k^{th} cluster centroid, and $\|c_k\|$ refers to norm of c_k.

Clusters are also identified using hierarchical clustering methods. Only the average linkage method and Ward's method performed well as other hierarchical methods

Table 6: Sector-wise list of the 161 KSE companies.

Sr. no.	Sector	Companies
S1	Oil and gas	BYCO, MARI, NRL, OGDC, POL, PRL, PSO, and SHEL
S2	Chemicals	CPL, DAWH, DSFL, DYNO, ENGRO, FFBL, FFC, ICI, LOTPTA, NICL, and SITC
S3	Forestry (paper and board)	CEPB
S4	Industrial metals and mining	HSPI and INIL
S5	Construction and materials (cement)	CHCC, DGKC, DNCC, EMCO, FCCL, FECTC, FRCL, GWLC, HADC, KOHC, LUCK, MLCF, and PIOC
S6	General industrials	GHGL, MACFL, PKGS, and SIEM
S7	Electronic and electrical goods	PCAL
S8	Engineering	AGTL, HINO, KSBP, MTL, and PECO
S9	Industrial transportation	PNSC
S10	Automobile and parts	ATBA, ATLH, BWHL, DFML, GHNL, GTYR, HCAR, INDU, PSMC, and SAZEW
S11	Beverages	MUREB
S12	Food producers	CHAS, CLOV, DWSM, HABSM, HAL, JDWS, MIRKS, MZSM, NATF, NESTLE, PNGRS, QUICE, SANSM, SGML, and SHSML
S13	Household goods	PAEL and SING
S14	Personal goods (textile)	ANL, BNWM, COTT, CRTM, GADT, GLPL, IBFL, IDRT, IDYM, KOIL, KTML, LOTPTA, MFTM, MUKT, NCL, NML, PSYL, RAVT, REWM, and SAIF
S15	Tobacco	PAKT
S16	Healthcare equipment and services	SHFA
S17	Pharma and biotech	FEROZ, GLAXO, HINOON, and SEARL
S18	Travel and leisure	PIAA
S19	Fixed line telecommunication	PTC
S20	Electricity	HUBC, JPGL, KESC, KOHE, KOHP, SEL, and SEPCO
S21	Commercial banks	BOP, ESBL, FABL, FDIBL, KASBB, MCB, MEBL, NBP, and NIB
S22	Nonlife insurance	AICL, CENI, EFUG, HICL, PAKRI, and RICL
S23	Life insurance	EFUL
S24	Financial services	FCSC, FECM, JOVC, JSCL, OLPL, and SIBL
S25	Equity investment instruments	FFLM, FHAM, FHBM, FPJM, FUDLM, GASF, FNBM, MODAM, PAKMI, PMI, and SCM
S26	Technology hardware and equipment	PTEC

Table 7: Validity indices for estimated mixing coefficients and ranked $\overline{R}^2$ clustering.

Clustering approach			Validity index (ranks)	
			Calinski–Harabasz	Davies–Bouldin
Proposed techniques	Estimated mixing coefficients		44.1597 (1)	4.0552 (1)
	Ranked $\overline{R}^2$	JADE	15.6017 (2)	4.1257 (2)
		FastICA	12 (3)	4.2017 (3)
		SOBI	11.5935 (4)	4.2707 (4)
Existing techniques	Average linkage method		10.2368 (6)	5.3113 (6)
	Ward's method		14.9963 (5)	5.2687 (5)

identified a large number of clusters consisting of one or two members. Hence the two indices are calculated for the clusters formed by the two proposed approaches and two existing approaches, i.e., the average linkage method and Ward's method to cluster variables. The results are presented in Table 7. The relative position of each of the indexes is presented in parentheses for quick comparison.

As depicted in Table 7, the performance of both proposed methods is better than the existing techniques. The performance of the estimated mixing coefficient approach is the best. As discussed earlier, the large value of the Calinski–Harabasz index and the small value of Davies–Bouldin are considered better. Both indices awarded rank 1 to the proposed estimated mixing coefficients approach. The performance of JADE is better for the ranked $\overline{R}^2$ approach. Results of FastICA with ranked $\overline{R}^2$ are on the third place, algorithm SOBI with the same approach are placed on number 4. Among the existing approaches, the results of Ward's method are better than the average linkage method.

5. Conclusion

In this paper, we presented two innovative approaches for clustering of multivariable datasets. The first approach is based upon the sum of squares of mixing coefficients, and the second is established using the ranking pattern of coefficient of determination of reconstructed and original series. Internal as well as external structure of clusters is explored. The compatibility of the clusters is contrasted with the available grouping mechanisms. It is concluded that the identification of clusters of stocks in their appropriate sectors is difficult because of the uncertain behavior of some stocks. Thus clustering should not be mixed up with predefined grouping whatsoever.

Gauging the cluster quality using Calinski–Harabasz index and Davies–Bouldin index, we conclude that the performance of both proposed techniques is better than the existing traditional techniques. Our evaluation indicates that the estimated mixing coefficients approach can be regarded as a better approach among the proposed techniques.

In the future, the current study can be extended to evaluate the performance of our suggested approaches on different types of datasets, e.g., biomedical, chemometrics, and signal processing datasets. Moreover, a criterion based on the level of independence of ICs may be explored for the identification of clusters. The proposed approach may also be explored for datasets having noise and outliers. The ICA-based identification of a cluster of observations may also be attempted.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Saima Afzal conceived and designed the study. Muhammad Mutahir Iqbal supervised the study and reviewed the manuscript. Ayesha Afzal did the computational work and wrote the manuscript. Hassan Bakouch helped in the computation and the analysis of results. Sadiah Aljeddani edited the manuscript and suggested a few areas for further study. All the authors discussed the results and contributed to the final manuscript.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant: Project Code (22UQU4310037DSR03).

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–A decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[2] A. Fahad, N. Alshatri, Z. Tari et al., "A survey of clustering algorithms for big data: taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267–279, 2014.

[3] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[4] I. R. Keck, E. W. Lang, S. Nassabay, and C. G. Puntonet, "Clustering of signals using incomplete independent component analysis," in *Computational Intelligence and Bioinspired Systems, 8th International Work-Conference on Artificial Neural Networks, IWANN Barcelona*, Spain*Paper presented at the*, Spain, 2005.

[5] B. B. Jamal and J. T. Kent, "Independent component analysis: an approach to clustering," in *International Conference on Modeling, Simulation & Visualization Methods, MSV Las Vegas Nevada*, USA, 2009.

[6] M. S. Islam, M. S. Islam, and M. Naseer, "PCA versus ICA in visualization of clusters," in *International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment*, Bangladesh, December 2012.

[7] M. S. Reza, M. Nasser, and M. Shahjaman, "An improved version of kurtosis measure and their application in ICA," *International Journal of Wireless Communication and Information Systems*, vol. 1, no. 1, 2011.

[8] F. R. Bach and M. I. Jordan, *Finding Clusters in Independent Component Analysis*, University of California at Berkeley, 2002.

[9] I. R. Keck, S. Nassabay, C. G. Puntonet, and E. W. Lang, "A new approach to clustering and object detection with independent component analysis," in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach*, J. Mira and J. R. Álvarez, Eds., pp. 558–566, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[10] I. R. Keck, F. Theis, P. J. Gruber et al., "Automated Clustering of ICA Results for fMRI Data Analysis," in *International Confer-ence on Computational Intelligence in Medicine and Healthcare (CIMED)*, Lisbon, Portugal, Paper presented at the 2nd 2005.

[11] E. H. C. Wu and P. L. H. Yu, "ICLUS: a robust and scalable clustering model for time series via independent component analysis," *International Journal of Systems Science*, vol. 37, no. 13, pp. 987–1001, 2006.

[12] C. J. Lu and C.-C. Chang, "A hybrid sales forecasting scheme by combining independent component analysis with *K*-means clustering and support vector regression," *The Scientific World Journal*, 2014.

[13] M. Azam and N. Bouguila, "Speaker classification via supervised hierarchical clustering using ICA mixture model," in *Image And Signal Processing: 7th International Conference, ICISP 2016, Trois-Rivières, QC, Canada, May 30 - June 1, 2016, Proceedings (193-202)*, A. Mansouri, F. Nouboud, A. Chalifour, D. Mammass, J. Meunier, and A. Elmoataz, Eds., Springer International Publishing, Cham, 2016.

[14] M. Nascimento, F. F. E. Silva, T. Sáfadi, A. C. C. Nascimento, T. E. M. Ferreira, L. Barroso et al., "Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data," *PLoS One*, vol. 12, no. 7, 2017.

[15] E. Gultepe and M. Makrehchi, "Improving clustering performance using independent component analysis and unsupervised feature learning," *Human-centric Computing and Information Science*, vol. 8, no. 25, 2018, https://doi.org/10.1186/s13673-018-0148-3.

[16] J. Durieux and T. F. Wilderjans, "Partitioning subjects based on high-dimensional fMRI data: comparison of several clustering methods and studying the influence of ICA data reduction in big data," *Behaviormetrika*, vol. 46, pp. 271–311, 2019, https://doi.org/10.1007/s41237-019-00086-4.

[17] K. Shahina and P. T. S. Kumar, "Similarity-based clustering and data aggregation with independent component analysis in wireless sensor networks," *Transactions on emerging telecommunication technologies*, vol. 33, no. 7, 2022, https://doi.org/10.1002/ett.4462.

[18] P. Boonyakitanont, B. Gabrielson, I. Belyaeva et al., "An ICA-based framework for joint analysis of cognitive scores and MEG event-related fields," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3594–3598, 2022.

[19] A. D. Back and S. A. Weigend, "A first application of independent component analysis to extracting structure from stock returns," *International Jornal of Neural Systems*, vol. 8, no. 4, pp. 473–484, 1997.

[20] E. G. Prieto, *Independent Component Analysis for Time Series*, Ph. D., Charles III University of Madrid, Spain, 2011.

[21] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Radar and Signal Processing IEE Proceedings F*, vol. 140, no. 6, pp. 362–370, 1993.

[22] A. Hyvärinen, "Fast and robust fixedpoint algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[23] A. Hyvärinen and E. Oja, "A fast fixed point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[24] A. Belouchrani, K. A. Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[25] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[26] S. Afzal and M. Iqbal, "A new way to order independent components," *Journal of Applied Statistics*, vol. 43, no. 9, 2016.

[27] S. Afzal, M. Iqbal, and A. Afzal, "On the number of independent components: an adjusted coefficient of determination based approach," *Electronic Journal of Applied Statistical Analysis*, vol. 14, no. 1, pp. 13–27, 2021.

[28] T. Wittman, *Time-series Clustering and Association Analysis of Financial Data*, University of Texas, Austin, 2002.

[29] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[30] J. A. Hartigan, *Clustering Algorithms*, Wiley, New York, 1975.

[31] N. Basalto and F. De Carlo, "Clustering financial time series," in *Practical Fruits of Econophysics*, H. Takayasu, Ed., Springer, Tokyo, 2006.

[32] R. B. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, 1974.

[33] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, *PAMI-*, 1979.